TURNER, DANA P., Ph.D. An Analysis of Rater Effects in Reviews of Scientific Manuscripts. (2017)
Directed by Dr. John T. Willse. 160 pp.

In the peer review process used by scientific journals, ratings of manuscripts are obtained and used to make publication decisions. Though concerns have been raised about reviews given to scientific manuscripts, little has been done to address the effects of reviewer severity bias on decision making. In other settings, the methods of Generalizability Theory and Many-Facet Rasch Measurement often have been used to investigate and address such effects. The purpose of this study is to use Generalizability Theory and Many-Facet Rasch Measurement to examine the effects of reviewer severity on the ratings and decisions made during the peer review of scientific manuscripts. The merits of each method and their utility in this novel context also are assessed.

Deidentified peer reviews ($N = 635$) that used a five-item rating scale were included in a two-facet, partially nested Generalizability Theory analysis and subsequent Decision Studies. Many-Facet Rasch Measurement analysis of the data produced reviewer severity measures and manuscript publishability measures corrected for reviewer severity. Multinomial logistic regression analysis was used to compare manuscript decision categories predicted by average raw scores and Many-Facet Rasch Measurement corrected scores. Reviewer severity rankings also were compared using raw and adjusted methods.

The results of the Generalizability Theory analysis revealed that reviewers nested within manuscripts account for 35.48% of the variance in publishability scores. Manuscripts accounted for 12.21% of the total variance, and items accounted for 15.22%

of the total variance. Decision Studies indicated that an unrealistic number of reviewers and items would be needed to increase the generalizability coefficient and index of dependability to acceptable levels and that other methods of improving reliability should be employed. When the average raw total score was used to predict manuscript decision category, the overall percentage of manuscripts that were correctly classified using the average raw total score was 55.15%. Using the manuscript publishability measure (theta), the percentage of manuscripts that were correctly classified when the publishability measure was used was 52.49%, suggesting differences in classification, if a manuscript publishability measures corrected for reviewer severity were used. The reviewers' average raw ratings and the reviewers' severity measures had a Spearman rank-order correlation of -0.6083, which demonstrates differences likely attributable to the adjustment for manuscript quality in the severity measure.

These findings indicate that reviewers are inconsistent in their reviews of manuscripts. Reviewer severity bias can be addressed with Many-Facet Rasch Measurement adjustments, but additional reviewer training may be needed to improve the reliability of manuscript scores. Both Generalizability Theory and Many-Facet Rasch Measurement contributed to the findings of the study and to understanding reviewer behavior. These methods show potential for increasing the capacity for more fair and accurate rating methods in the peer review of scientific manuscripts.

# AN ANALYSIS OF RATER EFFECTS IN REVIEWS OF SCIENTIFIC MANUSCRIPTS

by

Dana P. Turner

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2017

Approved by

_____
Committee Chair

APPROVAL PAGE

This dissertation written by DANA P. TURNER has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair _____

Committee Members _____

_____

_____

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

ACKNOWLEDGEMENTS

This dissertation represents the culmination of work that has involved the efforts of many individuals. I would like to thank my committee chair, Dr. John Willse, for his guidance throughout this process. His contribution of ideas and time has not only served to make this dissertation a reality but has increased my knowledge and understanding of the content area. I am grateful to my committee members Dr. Robert Henson, Dr. Richard Luecht, and Dr. Devdass Sunnassee for contributing their expertise and valuable ideas that have shaped this work. Their input has greatly enhanced my dissertation.

I would also like to thank the others who have influenced me along the way. In addition to the aforementioned professors, I would like to thank the other professors from the Educational Research Methodology department for their contributions to my education. The knowledge and skills I have gained in this program have directed my future plans and will be used throughout my career. Outside of the University of North Carolina at Greensboro, I have received influence and support from numerous people. Of these many influences, I must acknowledge my family who instilled in me a love of learning and reading at a young age, without which I would never have pursued a doctorate.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

# CHAPTER I

# INTRODUCTION

In many areas, quality or suitableness is judged by obtaining ratings from human judges. These ratings are then used in decision making, and the results are trusted to be accurate and fair. Although this process has been used throughout history, questions have been raised about the effects of raters who may possess characteristics that influence their ratings. Methods of exploring and accounting for these potential problems have been developed, including Generalizability Theory and Many-Facet Rasch Measurement, which have traditionally been used in the education field. One other important area where ratings are often used in decision making is the peer review process of scientific journals. Though this process is at the heart of scientific investigation, little has been done to assess the effects of rater severity bias in peer review. The purpose of this study is to investigate rater severity bias in reviews of scientific manuscripts using Generalizability Theory and Many-Facet Rasch Measurement. The utility of these two methods in this context are assessed, and the implications for policy and future research are explored.

## Statement of the Problem

When different people provide ratings, their results are likely to differ (de Gruijter, 1984). People may come from different backgrounds and have different experiences that may influence their ratings (Eckes, 2008; Eckes, 2009). Expectations

may be different, which can shape what these people view as excellent or poor. Additionally, in situations where scale items are used, raters may interpret these items differently (Hoyt, 2000). This variability associated with raters introduces construct-irrelevant variance into the score of the object of the measurement (Eckes, 2009).

Often called rater bias in many contexts, the problem has the potential to impact outcomes of rating situations. Rater bias can affect the mean, variance, or covariance of ratings (Hoyt, 2000). Additionally, the reliability of performance ratings is known to be low (Houston, Raymond, & Svec, 1991). However the possibility exists that, with multiple raters, the ratio of true score variance to error variance can be improved over that with just one rater (Wilson, 1988). One area where ratings from multiple individuals are used is the review of manuscripts submitted to scientific journals. While numerous ratings are regularly provided in this domain, little research has been conducted on the effects of rater harshness or leniency, described here as rater severity bias, in peer review of scientific manuscripts.

## Issues in Manuscript Review

Within groups of peer reviewers, some reviewers will consistently provide more positive reviews, and some reviewers will reliably provide negative reviews (Cicchetti & Conn, 1976; Raymond & Viswesvaran, 1993). The reasons for these differences may be numerous and complex but likely are related to experiences and personality. Peer reviewers often have diverse backgrounds and experiences (Rothwell & Martyn, 2000). They may have different degrees and training, and their beliefs about topics of research also vary and are shaped by their experiences. Within any field, there are topics that

generate controversy and issues that cannot be agreed upon. Reviewers' personal opinions on such topics likely influence their ratings of manuscripts (Rothwell & Martyn, 2000). If a manuscript expresses views that are in contrast to those of the reviewer, the reviewer may be unlikely to rate that work very highly unless the evidence is extremely convincing. Peer reviewers also differ in their training on the assessment of research quality. Some may be very experienced in critiquing scientific work, while others may be clinicians or practitioners with less research experience. Such individuals could produce very different reviews of the same work. While having variety in peer reviewers provides well-rounded reviews (Bornmann, Mutz, & Daniel, 2010), editors should be aware of the potential problems.

In deciding whether to publish submitted manuscripts, journal editors must rely on the opinions of the peer reviewers (Bornmann, et al., 2010; Rothwell & Martyn, 2000; van Rooyen, Black, & Godlee, 1999). They must have faith in the quality and fairness of these reviews to make decisions about the manuscripts. If very harsh reviewers rate a manuscript, it may be rejected when it would have been accepted had more lenient reviewers been involved. Journal editors must be conscious of consequences on the impact factor of the journal. Perhaps a good article that was rejected by harsh reviewers could have contributed to increasing the impact factor. On the other hand, accepting a weak manuscript reviewed by lenient reviewers could have the opposite effect. Understanding the tendencies of reviewers could be helpful in making such decisions.

While few studies have been conducted to address these concerns, the available research supports the lack of consistency among reviewers (Marsh & Ball, 1981). One

study evaluated the review of abstracts for a scientific meeting and found great variability among reviewer ratings (Cicchetti & Conn, 1976). Similarly, a study including both journal reviews and meeting abstracts found little agreement among reviewers (Rothwell & Martyn, 2000). In a stride toward improving the quality of the peer review process, a group of researchers developed an instrument to be used in assessing the quality of submitted peer reviews (van Rooyen, et al., 1999). While this instrument assesses whether and to what extent reviewers address important aspects of the manuscript, it does not address possible reviewer severity bias.

## Methods of Addressing Rater Effects

Existing research on the effects of raters in peer review has used correlation, factor analysis, and analysis of variance methods (Marsh & Ball, 1981). In other fields, additional methods of addressing rater effects have been developed. Multiple approaches have been used to address this problem, with many addressing special cases of rating situations (de Gruijter, 1984; Houston et al., 1991; Raymond, Harik, & Clauser, 2011; Raymond & Viswesvaran, 1993; Wang & Yao, 2013; Wilson, 1988). Two of the most widely used approaches are Generalizability Theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) and Many-Facet Rasch Measurement (Linacre, 1989). These two approaches are the focus of this research.

## Purpose of this Study

This study applies methods commonly used in performance assessment to the area of peer review and evaluates the potential of these methods for this area of research. The purpose of this study is to use Generalizability Theory and Many-Facet Rasch

Measurement to examine the effects of rater severity on the ratings and decisions made during the peer review of scientific manuscripts. These two methods have not previously been applied in this context, but their application in other contexts suggests potential utility in peer review. Potential changes to the reviewer rating system also are explored.

## Research Questions

### Generalizability Theory

Research Question 1: What proportion of variance in observed scores is attributable to reviewer variation, and how does this compare to the proportion of variance attributable to other sources?

Research Question 2: Do the results of a Generalizability Theory Decision Study suggest that the conditions of measurement (i.e., number of reviewers and number of items) for manuscript reviews be changed?

### Many-Facet Rasch Measurement

Research Question 3: Do raw publishability scores versus theta scores predict meaningfully different manuscript decision classifications?

Research Question 4: How closely do ranks of the severity measure from a Many-Facet Rasch Measurement analysis compare to ranks using average raw ratings from each reviewer?

## Need for the Study

Research on quality improvement in peer review is greatly needed (Jefferson et al., 2007). In the past few decades, there has been much discussion of this topic, but little has been done to address the issue. Many researchers report viewing reviewer bias as a problem (Resnik, Gutierrez-Ford, & Peddada, 2008); however, few empirical studies have been conducted to work toward a solution. The difficulty of assessing the quality of peer review has been acknowledged, but work in this area is necessary if the process is to be improved (Jefferson, Wager, Davidoff, 2002; Smith, 1994). Steps must be taken to preserve the integrity of this important step in the dissemination of science.

This research also addresses a need for knowledge of methods of addressing rater effects. While Generalizability Theory and Many-Facet Rasch Measurement methods have been used for this purpose, the two methods rarely have been applied to the same data (Kim & Wilson, 2009; MacMillan, 2000; Sudweeks, Reeve, & Bradshaw, 2005). This raises questions about the soundness of these methods and the accuracy of conclusions drawn from findings of such analyses. Further work is needed to compare findings of the methods and to assess their utility in different types of rater situations. The results from such work have wide applicability beyond the field of peer review, as other fields that use ratings as an assessment method could benefit from enhanced knowledge of rater severity bias and improved methods for addressing this problem.

## Definition of Terms

To facilitate understanding of the concepts in this study, defining the commonly used terms that are the focus of this research is necessary. In the context of this study, the terms "rater" and "reviewer" or "peer reviewer" are used interchangeably. These terms all refer to the individual who provides an evaluation of a scientific manuscript. A "reviewer" of a manuscript is considered to be a "rater." A "rater" is not considered to be a "reviewer" in all contexts and will not be described as such when studies in fields other than peer review are discussed. The terms "rating" and "review" are used interchangeably when referring to the evaluation that is provided by an individual in the context of peer review. A "review" is a "rating," but a "rating" is not considered a "review" outside the peer review context. "Rater bias" refers to differences in raters that may occur due to differing opinions, perspectives, and experiences or differences in interpretation of the rating scale items (Hoyt, 2000). "Rater severity" is a rater's tendency to provide a more negative rating than expected. "Rater leniency" refers to the tendency to provide a more positive rating than expected. "Rater severity bias" describes the more specific circumstances of raters affecting results because of their tendency to be harsh or lenient. Lastly, "publishability" describes the extent to which a scientific manuscript contains features that make it desirable for publication.

## Content Overview

This dissertation includes a review of the literature on peer review and its existing problems, Generalizability Theory, Many-Facet Rasch Measurement, studies using both Generalizability Theory and Many-Facet Rasch Measurement to address rater effects, and

validity considerations. Following is a description of the methods used in the study, including details on how each research question is addressed. The data used in the study is fully described, along with the instrument used in data collection, and procedures for data analysis are explained. Next, the results of each analysis method are described, and these results are applied to the four research questions. A discussion of the results follows, including conclusions, implications, and limitations of the study.

# CHAPTER II

# REVIEW OF LITERATURE

The following review of the literature begins with an overview of perceptions of the peer review process, followed by an appraisal of existing research in the field. The intricacies of Generalizability Theory and Many-Facet Rasch Measurement are described, and existing relevant research using these methods is examined. Finally, the role of validity in this research area is considered.

## Issues in Peer Review

The peer review system is an important part of the dissemination of scientific research and is in place to regulate the quality of published journal articles. Despite the implicit trust in this system to uphold the highest standards of science, the peer review process can only work if quality is maintained and unbiased evaluation of potential publications is performed (Grainger, 2007). Peer review is not limited to scientific journals, but is widely employed in other areas such as teaching ability assessment, clinical skills performance, and research grant applications (Lee, Sugimoto, Zhang & Cronin, 2013; Smith, 2006). Although the current study focused on the review of scientific manuscripts, all of these instances can be affected by failures in the review process.

Little research has been done on peer review and much of the evidence of problems is anecdotal (Lee et al., 2013; Lock, 1985). Many opinions exist but have not

been supported by empirical research. The existing body of literature has been judged as of limited quality with most of the studies focusing on the effects of blinding of reviewers and authors (Jefferson, Rudin, Brodney Folse, & Davidoff, 2007). This provides, at least, a starting point but is not enough to either justify or refute criticisms of the system.

## Perceptions of Peer Review

The peer review process is often criticized because of its potential to be unfair and produce conflicting or unexpected results (Rennie, 2003). Oftentimes, reviewers disagree in their views, and the disposition of manuscripts may not truly reflect the quality therein. The underlying factor affecting the system is the high level of human involvement (Rennie, 2003). Human beings introduce their personal preferences, prejudices, and even shortcomings into the process, producing less than perfect outcomes. Therefore, peer review is by nature inconsistent and even subjective (Smith, 2006). In fact, reviewers may not even carry out the same procedures when scrutinizing manuscripts, evidenced by the fact that only 14% of peer reviewers reported undergoing formal training in the process of reviewing a manuscript (Snell & Spencer, 2005). Additionally, some reviewers simply may not be motivated enough to provide a high quality review (Garcia, Rodriguez-Sanchez, & Fdez-Valdivia, 2015).

In a survey of peer review perceptions among researchers, bias was the second most commonly reported ethical problem, endorsed by 50.5% of respondents (Resnik, 2008). Principle investigators and post-doctoral fellows tended to report that bias was a problem more often than technicians and staff scientists, suggesting that more experience with the peer review process could lead to this attitude. The most commonly reported

problem was incompetent review (61.8%), which also was more commonly reported by principle investigators and post-doctoral fellows. These findings suggest that the peer review process is questioned even by those who participate as reviewers themselves.

## Rater Agreement

Lack of rater agreement was one of the first articulated concerns about peer review (Marsh & Ball, 1981; Lock, 1985). This has continued to be acknowledged as problematic and in need of attention (Suls & Martin, 2009). The issue persists in many topic areas and situations. Reviewer disagreement is not limited to full manuscripts, but also has been identified as a potential problem in the review of scientific meeting abstracts (Rubin, Redelmeier, Wu, & Steinberg, 1993). Interestingly, rater agreement in article recommendation (i.e., accept, reject, etc.) has been shown to be different in different scientific areas (Lock, 1985), suggesting the degree of the problem is not the same in all disciplines.

## Rater Bias

Rater bias can be understood as differences in the way criteria for evaluation are understood and applied (Lee et al., 2013). Like the evidence for most criticisms of peer review, the evidence for bias is conflicting (Smith, 2006). Authors seem to feel that reviewers are biased toward them, but empirical evidence to support or refute these claims is lacking (Lock, 1985).

Many ideas exist as to why bias may occur. Bias can be a function of reviewer characteristics such as gender or nationality or even the content of the study (Benos et al., 2006; Gilbert, Williams, & Lundberg, 1994; Langfeldt, 2006; Lee et al., 2013; Link,

1998; Mahoney, 1977; Opthof, Coronel, & Janse, 2002; Rowland, 2002). Some evidence exists that reviewers provide more favorable recommendations for manuscripts with positive outcomes (Emerson et al., 2010) and for meeting abstracts with positive outcomes (Callaham, Wears, Weber, Barton, & Young, 1998). In one study of scientific meeting abstracts, reviewers who had submitted abstracts themselves provided lower ratings of abstracts than reviewers who had not submitted their own abstracts (Blackburn & Hakel, 2006). As is apparent from these findings, many aspects seem capable of affecting the potential for bias. The process a reviewer goes through when evaluating a manuscript could provide valuable information about peer review but has not been investigated (Kassirer & Campion, 1994).

Some of the earliest work in the area of peer review examined reviewer agreement and "systematic response bias" in reviews of manuscripts submitted to the *Journal of Educational Psychology* (Marsh & Ball, 1981). Reviewers provided a recommendation for articles and rated the articles on four subscales of aspects of significance and quality. Correlations were computed for a multitrait-multimethod matrix, and principal components analysis was conducted to explore the potential for using a weighted average of reviewer scores to improve reliability. Bias was also explored by examining deviations from means, and this rater response bias was used to correct the original rating. The effects of this correction were explored through analysis of variance. The authors found that reviewers did not tend to agree on ratings they gave to individual items. The weighted average of the items improved reliability slightly. Analysis of variance findings showed significant differences among reviewers in their

total score ratings and in their recommendation for the manuscript, assuming there were no systematic differences in the quality of the manuscripts. Rater response bias accounted for 31% of total score differences and 27% of recommendation differences, but the effects were not statistically significant. Correcting for rater response bias did not produce a statistically significant effect or improve reliability.

Later research was conducted on reviews of manuscripts submitted to the same journal after the previous analysis (Marsh & Ball, 1989). In an attempt to expand upon previous results, similar methods were employed, but, in this instance, authors were asked to complete additional experimental items to rate various aspects of the manuscript. A factor analysis revealed four components present in this 21-item questionnaire. These four components corresponded to the four items that were originally used. Further analysis did not provide evidence that the use of these dimensions provided improved results over simply using the overall recommendation for the manuscript.

Siegelman (1991) classified reviewers of the journal *Radiology* as zealots, pushovers, mainstream, demoters, and assassins. Reviewers of this journal use a scale from one to nine to rate each manuscript. In this study of reviewer tendencies, mean ratings were calculated for each reviewer, and their deviations from the overall mean were used to classify them into one of the five aforementioned categories. Zealots and pushovers provided more favorable ratings and demoters and assassins provided less favorable ratings. After taking into account the merit of the manuscripts being reviewed, the author was able to conclude that there were many divergent reviewers in the pool,

suggesting the importance of accounting for such variation when assigning reviewers and making decisions about manuscripts.

In the past, steps were taken to prevent bias based on characteristics such as status, rank, institution, gender, and research point of view (Lock, 1985). Blinding of authors, reviewers, or both is an attempt at preventing some of these biases from affecting peer review. Blinding was introduced to prevent bias, but blinding also reduces the amount of information available to the reviewers when evaluating the manuscript, potentially decreasing the quality of the review (McNutt, Evans, Fletcher, & Fletcher, 1990). Another criticism of blinding is that it does not always work as designed because articles often contain clues to the identities of the authors (Lock, 1985). This is especially true for manuscripts written by well-known authors who, when their names were masked, were still identifiable (Justice et al., 1998).

As the sole widely-used defense again bias, blinding has produced mixed findings when its effects have been studied. Some findings have shown little or no difference in the quality of reviews for unblinded versus blinded groups (van Rooyen, Godlee, Evans, Smith, & Black, 1999). On the other hand, the opposite conclusion that blinded reviews are of higher quality has been found (McNutt et al., 1990). Another study found that blinding did not affect review quality but did reduce the likelihood that reviewers would recommend manuscripts be rejected (Godlee, Gale, & Martyn, 1998). Still another found that there were no differences in the quality of reviews between reviewers who were identified and reviewers who were anonymous or their recommendations for the manuscript (van Rooyen, Godlee, Evans, Black, & Smith, 1999). Despite the conflicting

evidence, many researchers still prefer blinded reviews. One study found that 56% of researchers report preferring double-blind review, and 25% prefer single-blind review (Ware, 2008). While blinding might be preferred, the evidence suggests that it alone is not enough to prevent bias in manuscript review.

## Work in Other Contexts

Outside the realm of peer review, more attention has been given to rater effects. Those in the education and performance assessment fields have been aware of these concerns for many decades (Dunbar, Koretz, & Hoover, 1991; Raymond & Houston, 1990; Wolfe, 2004). In classroom assessments, as well as broader-scale testing programs, rating systems have regularly been employed, and methods of controlling their quality have been explored.

Language testing is one particular area where influences on raters have been extensively studied and variability in rater behavior has been examined (Kondo-Brown, 2002; Lumley & McNamara, 1995; Johnson & Lim, 2009). Additionally, employment testing has long been in need of methods to address rater effects (Holzbach, 1978). The nature of both language assessment and workplace performance assessment allows for the introduction of sources of variability and bias, making measurement more challenging.

Investigations into the introduction of bias into such rating situations have revealed complex reasons for this problem. Evidence exists that raters may unconsciously form judgments about ratees that affect their ability to provide objective ratings (Gingerich, Regehr, & Eva, 2011). Rater personality and the social context of the

situation also affect ratings (Yun, Donahue, Dudley, & McFarland, 2005), creating circumstances that are difficult to control. In fact, work in testing has found that rater severity may even differ across time (Congdon & McQueen, 2000), although contrary findings have occurred in workplace performance assessment (Kane, Bernardin, Villanova, & Peyrefitte, 1995). Taken together, the evidence for rater bias in these fields suggests that similar problems may exist in scientific peer review.

## Methods of Assessing Rater Effects

Rater effects may be assessed and accounted for in a number of ways. While some early work focused on ordinary least squares and weighted least squares methods (Raymond & Houston, 1990), two of the most often used and the most strongly supported methods are Generalizability Theory and Many-Facet Rasch Measurement. These methods take different analytical approaches and produce different types of results. These two methods are the focus of this research and are detailed in the following sections.

## Generalizability Theory

While Generalizability Theory was developed decades ago (Cronbach, et al., 1972), this method is still often used to study sources of variability in measurement (Lakes & Hoyt, 2008; Lin, 2014) and is considered highly applicable to rating occasions such as performance assessment (Brennan, 2000). Generalizability Theory is related to the dependability of measurements, which is the accuracy of generalizing from an observed score to a person's average of scores across all possible testing occasions (Brennan, 1983; Brennan, 2001; Cronbach, et al., 1972; Shavelson & Webb, 1991). A

measurement is considered to be a sample from the universe of admissible observations, which is all observations that can be treated as interchangeable in decision making. In Generalizability Theory, sources of measurement error are called facets (Brennan, 1983; Brennan, 2001; Cronbach, et al., 1972; Shavelson & Webb, 1991). Common examples of facets are items, occasions, and raters. Multiple facets can be included in one study. Because they are not sources of measurement error, persons are not considered facets in Generalizability Theory.

Generalizability Theory allows multiple sources of error to be considered at once (Brennan, 1983; Brennan, 2001; Cronbach, et al., 1972). Under Generalizability Theory, the error term from classical test theory (Lord & Novick, 1968) is divided among facets (e.g., raters), or systematic sources of variability, and random error. A variance component is estimated for each facet, interactions of facets, and a residual. Facets are considered to be random when the sample is exchangeable with another sample from the same universe. If the number of conditions of a facet is the same as the number in the universe, the facet is considered fixed (Cronbach, et al., 1972; Shavelson & Webb, 1991).

**Study Designs**

Many design possibilities exist for Generalizability Theory studies. In a crossed design, all conditions of a facet occur with all conditions of the other sources of variability. This would occur in a study where each rater rated each person using all available items. In nested designs, all conditions of a facet do not appear with all conditions of another source of variability (Cronbach et al., 1972; Shavelson & Webb, 1991). For example, all raters may not rate every person but may rate subsets of the

available people; the rater facet is nested within the person facet. Partially nested designs include both crossed and nested facets. In this design, rater could be nested within person and crossed with items (i.e., raters only rated some people, but all items were administered to all people by all raters). Balanced designs have no missing data and have equal sample sizes for all levels of nested facets (Brennan, 2001). In unbalanced designs, the sample size differs for levels of nested facets. If raters are nested within persons, the number of raters may be different for different people.

**Notation**

Common notation for Generalizability Theory studies includes each source of variability and each variance component. For example, a crossed study including persons, raters, and items as sources of variability is represented by Equation 1 and includes the following notation:

Persons ($p$): $\sigma_p^2$

Raters ($r$): $\sigma_r^2$

Items ($i$): $\sigma_i^2$

Person x Rater interaction ($p$ x $r$): $\sigma_{pr}^2$

Person x Item interaction ($p$ x $i$): $\sigma_{pi}^2$

Rater x Item interaction ($r$ x $i$): $\sigma_{ri}^2$

Residual of unique combinations of Persons, Raters, Items, unmeasured facets, and random error ($p$ x $r$ x $i$, $e$): $\sigma_{pri,e}^2$

Observed score ($x$): $\sigma_x^2$

$$\sigma_x^2 = \sigma_p^2 + \sigma_r^2 + \sigma_i^2 + \sigma_{pr}^2 + \sigma_{pi}^2 + \sigma_{ri}^2 + \sigma_{pri,e}^2 \quad (1)$$

The results of the analysis are the estimates of these variance components. When focusing on rater effects, the result of interest is the rater component ($\sigma_r^2$), which indicates the rater mean score variance across persons and items (Brennen & Johnson, 1995). The person x rater interaction variance component ($\sigma_{pr}^2$) indicates differences in how raters rank order persons, and the rater x item interaction variance component indicates differences in how raters order the difficulty of items.

A partially nested study that includes facets for item and rater nested with person would be notated as *(r:p)* x *i*. Rater nested within person would be represented as *r:p* with a variance component of $\sigma_{r,pr}^2$, item would be represented as *i* with a variance component of $\sigma_i^2$, and the item-by-rater interaction confounded with the three-way person-by-rater-by-item interaction and other sources of error would have variance component $\sigma_{ir,pri,e}^2$.

**Decision Studies**

A decision study uses the information from the generalizability study to design the best measurement conditions for the intended purpose. These results can provide the number of conditions of a facet needed to achieve the desired reliability (Cronbach, et al.,1972; Shavelson & Webb, 1991). Estimated variance components are calculated under different conditions of the facet, providing information about the conditions of a facet that would minimize error. For example, the optimal number of raters can be determined in a study that includes a rater variance component. Relative decisions

involve variance components that affect the ranking of the object of measurement (interactions between facets and the object of measurement). Absolute decisions are about only one object of measurement (e.g., person) not compared to the others and involve variance components for all interactions and main effects of facets. Relative error variance ($\sigma^2_{Rel}$; Equation 2) is used to calculate the generalizability coefficient ($\rho^2$; Equation 3), which is similar to the reliability coefficient and is used for relative decisions (Cronbach, et al.,1972; Shavelson & Webb, 1991).

$$\sigma^2_{Rel} = \sigma^2_{pr} + \sigma^2_{pi} + \sigma^2_{pri,e} \tag{2}$$

$$\rho^2 = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{Rel}} \tag{3}$$

Absolute error variance ($\sigma^2_{Abs}$; Equation 4) is used to calculate the index of dependability ($\Phi$; Equation 5), which is used for absolute decisions using.

$$\sigma^2_{Abs} = \sigma^2_r + \sigma^2_i + \sigma^2_{pr} + \sigma^2_{pi} + \sigma^2_{ri} + \sigma^2_{pri,e} \tag{4}$$

$$\Phi = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{Abs}} \tag{5}$$

The generalizability coefficient and index of dependability are typically interpreted similarly to Cronbach's alpha, the analogous concept of reliability used in Classical Test Theory (Brennan & Kane, 1977; Webb, Shavelson, & Haertel, 2006). Decision Studies offer different conditions of a facet (e.g., number of raters) and their corresponding generalizability coefficients and indexes of dependability. In applied

studies, coefficients of 0.80 have been considered to have reasonable reliability, while coefficients below 0.60 indicate poor reliability (Colliver, Verhulst, Williams, & Norcini, 1989; Shavelson & Webb, 1991). Coefficients above 0.80 reflect good reliability. Results from Decision studies can be compared to these criteria of acceptability to determine the conditions of facets that will raise reliability of the measurement situation to a desirable level.

**Limitations and Complexities in Generalizability Theory Analyses**

Generalizability Theory analyses function extremely well for fully crossed data, but deviating data structures can introduce complexities into to the analysis and limit results and their interpretations. Nested data limits the information that can be obtained from a Generalizability Theory study (Shavelson & Webb, 1991). Nested facets are confounded with error, and their variance components cannot be estimated separately. In a study with the rater facet nested within the person facet, a variance component for rater alone cannot be obtained. Therefore, the rater variance cannot be interpreted outside of the person variance. Variance components for interactions of the nested facet with other facets also cannot be obtained. For example, an item-by-rater interaction variance component cannot be obtained in the above mentioned study design. Only the variance component for the three-way interaction between manuscripts, item, and reviewer plus remaining sources of systematic and unsystematic variation not measured in the study can be obtained. This limitation complicates interpretation of the variance in scores that can be attributed to raters (or any other nested facet) and restricts understanding of potential interaction effects among facets.

The application of Generalizability Theory to unbalanced mixed effects study designs is complex but can be accomplished (Brennan, 2001; Luecht, 1989). When applied to balanced data, Generalizability Theory methods function relatively smoothly as analysis of variance (ANOVA) procedures. However, in real-world situations, data often is unbalanced. With data of this nature, variance components can be estimated in a variety of ways, requiring a complex choice among estimators (Brennan, 2001; Luecht, 1989). The analogous-ANOVA procedure, espoused by Robert Brennan (2001), results in unbiased estimates of variance components. Briefly, the procedure decomposes the total sums of squares in a way that is analogous to the manner of decomposition used in a balanced design. This method is appropriate for nested and partially nested designs and can be applied to crossed designs with missing data.

The results of Generalizability Theory studies conducted using analogous-ANOVA procedures are universe estimates and are not dependent upon missing data in the original dataset (Brennan, 2001). Similarly, Decision Study universe score variance results are not affected by missing data. However, error variance is affected by an unbalanced design. When a Generalizability Study is unbalanced, a Decision Study can be conducted with a balanced design using results from the unbalanced Generalizability Theory study, or a Decision Study can be conducted with an unbalanced design (Brennan, 2001). If the unbalanced design is the same as the Generalizability Theory study, the unbalanced nature of the Generalizability Study will affect the variance components. If the Decision Study uses an unbalanced design different from the Generalizability Theory Study, the unbalanced nature of the Decision Study will affect

the variance components. Often, formulas for obtaining Decision Study results from unbalanced data are complex.

## Many-Facet Rasch Measurement

Like Generalizability Theory, Many-Facet Rasch Measurement involves facets, but they are defined as elements of the measurement situation that have a systematic influence on scores (Engelhard, 1992; Linacre, Engelhard, Tatum, & Myford, 1994). These models are extensions of the Rasch model that are able to include additional variables beyond items and people (Engelhard 1992; Eckes, 2009; Engelhard, 2013; Wilson & Case, 2000) and have been applied in many fields, especially in performance assessment (Eckes, 2008; Engelhard, 1994; Farrokhi, Esfandiari, & Schaefer, 2012; Lunz, & Stahl, 1993; Prieto & Nieto, 2014; Wind, Engelhard, & Wesolowski, 2016). From these models, a proficiency measure, along with a standard error, can be obtained that is independent of the raters who provided the rating. In this way, the analysis corrects for error associated with rater severity. A fair score also can be obtained that provides the score that would have come from a rater with average severity. All estimates, including facets such as rater severity are on the same scale.

Many-Facet Rasch Measurement provides much information on the raters, including group-level and individual-level effects (Myford & Wolfe, 2003). Raters are ranked on severity and assigned a severity measure (Du & Brown, 2000; Engelhard, 2013). Information on the extent to which raters used the scale categories is also available, and the degree to which raters provide unexpected ratings can be determined. Another useful result is the fair average, which is the rater's mean rating adjusted for

examinee proficiency. This aids in interpreting severity by removing the effect of a particular rater's pool of examinees (Eckes, 2009). Additionally, rater separation statistics are available to summarize variability in the distribution of rater severity.

While the Many-Facet Rasch Measurement model is a main effects model, differential facet functioning analyses are possible (Du, Wright, & Brown, 1996). Information that can be obtained from these results includes whether raters rate each person in a similar manner regardless of the ratee's characteristics or whether raters with certain characteristics provide different ratings than raters with other characteristics. An additional use of these types of models is the measurement of rater accuracy (Engelhard, 1996). Because Many-Facet Rasch Measurement models are robust to designs that are not perfectly crossed, they can be used in numerous situations (Eckes, 2009). Two statistics are available for assessing reliability. The reliability of separation index estimates the ratio of true score to observed score variance, and the separation ratio provides the spread of measures of rater severity compared to the precision of the measures (Eckes, 2009).

Many-Facet Rasch Measurement models often are applied to the rating scale (Andrich, 1978) or partial credit models (Masters, 1982), which are extensions of the Rasch model that allow multiple response options for items. The rating scale model assumes that all items have the same structure to their rating scales. The partial credit model allows each item to have its own rating scale structure. This model may provide better fit over the rating scale model but introduces instability (Linacre, 2000; Wright, 1998).

**Notation**

The mathematical model includes terms for multiple aspects of the rating situation. For example, a model involving rater severity would include the terms shown in Equation 6.

$$ln\left[\frac{P_{\text{nijk}}}{P_{\text{nijk}-1}}\right] = \text{B}_\text{n} - \text{D}_\text{i} - \text{C}_\text{j} - \text{F}_\text{k} \quad (6)$$

$P_{\text{nijk}}$ = probability of person *n* receiving a rating of *k* on task *i* by rater *j*

$P_{\text{nijk-1}}$ = probability of person *n* receiving a rating of *k-1* on task *i* by rater *j*

$\text{B}_\text{n}$ = level/ability of person *n*

$\text{D}_\text{i}$ = difficulty of task *i*

$\text{C}_\text{j}$ = severity of rater *j*

$\text{F}_\text{k}$ = difficulty of receiving rating *k* relative to rating *k*-1

***Facets* Program**

The *Facets* software package computes Many-Facet Rasch Measurement models (Linacre, 2014). The program has been used extensively in many fields and is regularly updated. The results described above can be obtained from the program in the several tables and figures available as output (Myford & Wolfe, 2004). One such figure is a variable map that displays the logit scale, the ordered performance of each person, measures of trait difficulty, ordered rater severity measures, and thresholds of likelihoods for receiving each rating. Rating scale probability curves and fit indices also are available.

**Sample Size Considerations**

Sample size requirements for Many-Facet Rasch Measurement follow the general guidelines for Rasch analysis, which are designed to generate stability of measurement. This is an important consideration because smaller sample sizes can lead to poorer precision of estimates, decrease model fit, and lower robustness of estimates (Linacre, 1994). For studies using dichotomous items, a sample size of 30 is appropriate (Linacre, 1994; Wright & Tennant, 1996). In studies with polytomous items, a minimum sample size of 50 is needed. These sample sizes will provide calibrations that are stable within one logit. In high stakes situations, a minimum sample of 250 provides the needed increased stability.

**Model Fit**

Many-Facet Rasch Measurement analyses provide fit indices for each unit of each facet. These indices demonstrate the level to which the observed data matches the results from the model (Linacre, 2002; Eckes, 2009). Rater fit statistics provide the degree to which the raters used the rating scale consistently across persons. This also can be described as the level at which a rater provides unexpected ratings. Rater outfit is an unweighted mean-square statistic that is sensitive to very unexpected ratings from raters who are mostly consistent. An infit statistic is a weighted fit statistic that is sensitive to the occurrence of many unexpected ratings. Both outfit and infit statistics can range from zero to infinity and have expected values of one (Linacre, 2002; Myford & Wolfe, 2004). Values of greater than one indicate that raters exhibited more variability in their ratings than expected, and values less than one indicate less variability than expected. Person

infit and outfit statistics can be computed to examine how much a person's performance deviated from expected. Item infit and outfit statistics can be similarly computed and provide a measure of whether an item's difficulty was different than expected. Problematic values of fit statistics may signal problems with the model and should be reviewed and considered in interpretation of results. Fit statistics within a range of 0.50 to 1.50 typically are acceptable in the analysis (Linacre, 2002).

**Limitations in Many-Facet Rasch Measurement**

Many-Facet Rasch Measurement has many strengths and is able to provide considerable information in the study of rater effects (Eckes, 2009). Because this is a main effects model, no assessment of interactions between facets is conducted, which could be viewed as a weakness. In this method of analysis, the variability among raters is seen only through the single rater effect. Any interactions of the rater facet with other facets such as item or person cannot be captured in the analysis but is, instead, considered to be random error (Linacre, 1993). For example, an interaction between rater and item suggests that rater severity differed across items, but this cannot be captured in a Many-Facet Rasch Measurement analysis. Because of the method's deficiency, some raters could have the same level of rater severity, but their rating behavior could have had different patterns. If such problems exist, the model may have poor rater fit and potentially create difficulty in interpreting results.

## Comparison of Methods

Generalizability Theory and Many-Facet Rasch Measurement both provide methods of assessing and understanding the influence of raters. They each give insight into the circumstances surrounding the measurements of interest and offer methods of addressing potential measurement issues. However, these methods do not produce the same results and cannot be used interchangeably.

Generalizability Theory allows variance in ratings to be partitioned into several sources of variability (Brennan, 1983; Brennan, 2001; Cronbach, et al., 1972; Shavelson & Webb, 1991). The main purpose behind this technique is to assess the consistency of ratings, which can be understood as a type of reliability. Generalizability Theory expands upon Classical Test Theory methods by separating error that would have been indistinguishable into multiple sources. Through such an analysis, it becomes possible to understand the error variance associated with raters, as well as other aspects of the measurement situation. This makes it possible to know how well the universe score is predicted for examinees and how many raters would be necessary to minimize error associated with raters. The generalizability coefficient and the index of dependability give estimates of reliability that can be used in making decisions about the best measurement conditions. Overall, the goal of a Generalizability Theory analysis is to gauge how reliably the scores can be used for making generalizations about the object of measurement.

Many-Facet Rasch Measurement is based on a latent trait modeling the probability of a response (Engelhard, 1992; Eckes, 2009; Engelhard, 2013; Wilson &

Case, 2000). With this modeling approach, gaining considerable information about the raters is possible. Raters can be ranked in severity, and the degree to which their behavior deviates from expectation can be determined. This then allows for adjustment for rater influence, making it possible to examine performance after correcting for error associated with rater severity. Distributions of persons and items are available in addition to rater information, and reliability can be measured with the reliability of separation index and the separation ratio. Fit statistics demonstrate how well the observed performance of each unit of each facet matches that which would be expected from the model. Overall, a Many-Facet Rasch Measurement analysis focuses on adjusting for rater severity in the estimation of a latent trait and provides metrics for understanding and evaluating the estimated model.

Generalizability Theory and Many-Facet Rasch Measurement each address rater severity in different ways. The two methods do not accomplish the same thing, but they each offer an approach to addressing the problem. Generalizability Theory is an approach based on group behavior, while Many-Facet Rasch Measurement is individualized and can produce information about each rater (Myford & Wolfe, 2003). Many-Facet Rasch Measurement also uses this individual data to correct for error associated with rater severity. Generalizability Theory does not employ a correction but provides some guidance for optimal measurement situations through Decision Studies. In this case, error is dealt with by increasing the number of raters in the study. The interaction variance components found in Generalizability Theory are not found in Many-Facet Rasch Measurement. Instead, interactions are considered random error (Linacre,

1993). This conceptualization avoids characterizing unstable interaction variance as consistent, but it loses potential interaction effects. These effects may then appear as misfit in fit analyses. For example, if a Generalizability Theory analysis reveals large interaction effects, the fit of the model in a Many-Facet Rasch Measurement analysis of the same data is likely to be less than desirable. Primarily, Generalizability Theory is focused on reliability of ratings. Many-Facet Rasch Measurement provides measures of reliability, but evidence suggests these may overestimate the true reliability (Wilson & Hoskens, 2001). However, a large rater variance component from a Generalizability Theory analysis is expected to correspond to a Many-Facet Rasch Measurement analysis that finds raters to be reliably different in their ratings. While these two methods are not interchangeable, they do supply complementary information. In some circumstances, one approach may be more appropriate than the other, but these methods may be best used together.

## Studies Comparing these Methods

Generalizability Theory and Many-Facet Rasch Measurement have been used together in previous studies of rater severity (Kim & Wilson, 2009; MacMillan, 2000; Sudweeks et al., 2005). In these studies, the two methods have been used to supplement each other, but their results have also been compared. Such analyses demonstrate the potential for using Generalizability Theory and Many-Facet Rasch Measurement together to examine rating data.

One study used an English examination taken by 4,930 students with three raters who provided ratings on nine scales (MacMillan, 2000). In this example involving a true

to life case of large, sparse datasets, variance components calculated in the Generalizability Theory analysis and a histogram and logit range from the Many-Facet Rasch Measurement analysis were similar, and the authors concluded that raters did not vary considerably when analyzed with either method. However, Many-Facet Rasch Measurement did show more variability among raters than the variance component from Generalizability Theory analysis.

Another study involving 48 college undergraduate essays read by nine raters produced similar results with regard to variability (Sudweeks, et al., 2005). The facets that produced the highest variability were similar in both analyses (i.e., variance components for Generalizability Theory and separation index and separation reliability for Many-Facet Rasch Measurement). From these results, it appears that the two methods performed similarly in this instance.

Another such study used ratings of compositions from 229 high school sophomores. The Generalizability Theory variance component results revealed very small differences in rater severity, and Many-Facet Rasch Measurement rater severity measures provided similar conclusions (Kim & Wilson, 2009). The authors concluded that, while the two methods do not produce the same types of results, they each have their advantages, and the choice of which one to use could affect the conclusions of a study.

From these studies, it appears that Generalizability Theory and Many-Facet Rasch Measurement methods produce somewhat differing results but may provide similar conclusions. However, the sample sizes of these studies vary greatly, and other circumstances of measurement are not completely comparable. The use of both methods

was valuable in confirming or questioning the results of the studies. While these studies are good examples of how Generalizability Theory and Many-Facet Rasch Measurement can be used together, there is room for more detailed comparisons of results and discussions of relative advantages. Such descriptions could provide a deeper understanding of the two methods and their uses in studies involving raters, especially concerning error, reliability, and the effects of associated results.

## Application of Methods to Peer Review

When applying Generalizability Theory and Many-Facet Rasch Measurement to the field of peer review, the elements of each method must be reconceptualized. Because the two methods have traditionally been used in the education field, many of the commonly used descriptors refer to concepts such as ability and test-takers. In order to fully describe the potential for the methods to be used in peer review, terms germane to that context should be employed.

### Generalizability Theory

When using Generalizability Theory, the object of measurement will not be a person, but it will be a manuscript. This distinction is necessary to understand that manuscripts will be receiving ratings, not people. As with other analyses, items also will be a facet in the peer review context. The reviewers of manuscripts act as the raters of the manuscripts. Therefore, the sources of variation in such a design will be manuscripts, items, and reviewers. Each of these sources of variation will have a corresponding variance component. These will include a manuscript variance component ($\sigma_m^2$), an item variance component ($\sigma_i^2$), a reviewer variance component ($\sigma_r^2$), a variance component for

the manuscript-by-item interaction ($\sigma^2_{mi}$), a variance component for the manuscript-by-reviewer interaction ($\sigma^2_{mr}$), a variance component for the item-by-rater interaction ($\sigma^2_{ir}$), and a residual of unique combinations of manuscripts, reviewers, items, unmeasured facets, and random error $\sigma^2_{pri,e}$. More complex study designs may have additional facets and nesting within facets.

In a study of rater severity bias, the focus of the results will be on the reviewer variance component. This component of the results will provide the proportion of variance in observed scores that is attributable to reviewer variation. The manuscript-by-reviewer interaction variance component also is of interest and indicates differences in how reviewers rank order manuscripts. Additionally, the reviewer-by-item interaction variance component indicates differences in how reviewers order the difficulty of items. If the reviewer facet is nested within the manuscript facet, the reviewer variance component cannot be obtained independent of manuscript, and interactions of the reviewer facet with other facets (e.g., item) cannot be determined.

Findings from Decision Studies can assist in examining changes to the measurement conditions that would minimize error. Changes to the number of reviewers or the number of rating items can increase reliability to a more acceptable level. Therefore, the results of a Decision Study provide a useful guide for implementing changes to the manuscript review process.

**Many-Facet Rasch Measurement**

In the application of Many-Facet Rasch Measurement, the usual person facet will be the manuscript. The manuscripts will receive ratings by the reviewers. While the ability or proficiency of a person is estimated in most Many-Facet Rasch Measurement studies, the publishability of a manuscript is estimated in this context ($B_n$). The task represented in this situation is a manuscript receiving a score on the publishability items ($D_i$). Reviewer severity will serve as rater severity ($C_j$). Additional facets may be included in more complex studies.

In a study of rater severity bias, an important result is the manuscript scores corrected for reviewer severity. This will provide corrected ratings that are not affected by reviewer severity bias. Fair scores that represent the score that would have come from a reviewer with average severity also are useful. Information is available about the reviewers such as the extent to which reviewers used the scale categories, the degree to which reviewers provided unexpected ratings, and a fair average, the reviewers' mean ratings adjusted for manuscript publishability. Fit statistics will provide an assessment of whether the observed performance of each reviewer matches what would be expected from the model. Additional fit statistics for manuscripts will describe whether the observed publishability of each manuscript matches the expected publishability, and item fit statistics will detail the performance of manuscript rating items.

The results of these analyses will be improved manuscript publishability scores that are corrected for the effects of reviewer severity. These improved scores then can assist editors in making more informed decisions regarding manuscript publication.

With the newly corrected scores, manuscripts can be judged on their publishability without the influences of reviewer severity.

## Contribution to the Literature

Generalizability Theory and Many-Facet Rasch Measurement have not previously been applied to the field of peer review. However, there appears to be potential for such analyses to provide useful information to scientific journal editorial staff and to the research community. The publication process relies on reviews of manuscripts provided by peer reviewers, but this process provides little control to the journal staff, creating a need for improved methods (Bornmann, et al., 2010; Rothwell & Martyn, 2000; van Rooyen, Black, & Godlee, 1999). Currently, the effects of reviewer variability on manuscript ratings and the potential impact of adjustment for reviewer severity on manuscript ratings and decisions are not known.

Applying the methods of Generalizability Theory and Many-Facet Rasch Measurement in a new context also serves to test their applicability in other fields. In a peer review setting, data may not be in the pristine condition that is often the case when methods are first being developed and tested. Because peer review is voluntary, the process of finding enough individuals to complete the task can be complicated. Some manuscripts will have different numbers of reviewers because of availability or expertise concerns that are difficult to control. The connectivity of reviewers throughout a dataset may be less than desirable in this setting. For these reasons, the use of Generalizability Theory and Many-Facet Rasch Measurement many create a complex situation but one that is useful to study.

Few studies have incorporated both Generalizability Theory and Many-Facet Rasch Measurement into their methods. Those that have utilized the two methods found comparable but different results from the two analyses (Kim & Wilson, 2009; MacMillan, 2000; Sudweeks et al., 2005). These few studies suggest the need for more investigation into the performance of these methods. Because the methods accomplish related but different goals, they may be best used together to provide multiple types of results for rater studies. Such studies will expand knowledge on these methods themselves in addition to increasing understanding of rater behavior.

## Validity Considerations

This research has an important relationship to the concept of validity, especially the Generalizability Theory analyses. Generalizability is a necessary but not sufficient condition for validity (Kane, 1999; Kane, 2013). Validity can be defined as the accuracy of inferences made about the value of an attribute from an observed score and is characterized by degree as opposed to a yes or no judgment. As modernly defined, the degree of validity provides support for the interpretation and use of findings. When applying validity considerations to peer review, validity evidence serves to provide support for decisions made about manuscripts and whether they are publishable or not publishable.

Validity has long been discussed in other fields, especially educational and psychological testing. The early conceptualization of validity involved criterion-related (predictive and concurrent), content-related, and construct-related validity evidence Cronbach & Meehl, 1955). These three concepts relied on three different methods of

obtaining validity evidence, with criterion-related and construct-related employing empirical methods of assessment such as comparison to a different assessment and inter-item correlations, respectively.

Samuel Messick (1989) reconceptualized validity as a judgment of the degree that the theoretical basis and empirical evidence support inferences and decisions made based on assessment scores. This view unifies validity and brings theory into the forefront of decision making. Messick also incorporated generalizability into his understanding of validity as one aspect of validity evidence. In this context, generalizability of score interpretations refers to the extent to which the interpretations can generalize to different groups, settings, times, or tasks.

Michael Kane (2006) further expanded understanding of validity by presenting an argument-based approach, requiring both an interpretive argument and a validity argument. This framework requires the intended interpretations and uses to be made clear and supported by evidence. Further, generalizability evidence is considered necessary for establishing validity (Kane, 2013). Kane has conceptualized the universe of generalization as a universe of validity (1982). He suggests that a procedure for measurement can be called valid for the attribute of interest to the degree that the method is able to accurately estimate the expected value of the attribute across the universe of allowable observations. The results of a Generalizability Theory study can then be used to make a claim about a score over the universe of generalization (Kane, 2013).

In Generalizability Theory, large variance components can indicate inconsistencies within constructs. If constructs are not consistent, validity claims cannot

hold. If evidence for validity is weak, the measurement method is not an appropriate means on which to base decisions (Kane, 2013). From an associated Decision Study, a squared disattenuated validity coefficient can be obtained that represents the squared correlation between scores from a measurement procedure with perfect reliability and universe scores for the universe of generalization (Cronbach, et al.,1972; Shavelson & Webb, 1991). The squared disattenuated validity coefficient and the reliability coefficient comprise the dependability of inferences made from observed scores. In this way, not only is Generalizability Theory able to provide information about sources of variability in measurement, the method also is useful in obtaining a validity estimate for the measurement procedure.

In peer review, decisions must be made about whether to publish submitted scientific manuscripts. The criteria used in making decisions about the publishability of manuscripts should be supported by validity evidence. Generalizability Theory analyses can provide a method for obtaining such information. Gaining understanding of measurement methods for obtaining a manuscript's publishability will support or refute evidence for validity. An additional method of establishing validity evidence in peer review would be to calculate a validity coefficient using two measures of manuscript publishability, the reviewer's publication recommendation and the editor's publication decision. After Many-Facet Rasch Measurement analysis, these decision categories can be correlated with the publishability scores adjusted for reviewer severity. These methods of establishing validity evidence are important for determining whether decisions should be made based on the data from the measurements in question. If

appropriate validity evidence cannot be established, decisions about whether to publish a manuscript should not be made based on these measurements.

# CHAPTER III

# METHODS

This study assessed rater effects in reviews of scientific manuscripts. Generalizability Theory and Many-Facet Rasch Measurement methods were used to analyze the data. The data and analysis methods are described herein.

## Data

The data used in this analysis consisted of deidentified peer reviews of manuscripts submitted to a medical specialty journal. The data spans the time period from Fall 2013, the time when the manuscript rating system was implemented, to Fall 2015 and contains 918 reviews of 338 manuscripts. Individual reviewers contributed a mean of 2.2 reviews each. Of these reviews, 645 reviews of 311 manuscripts were completed by reviewers who contributed two or more (mean = 4.5) reviews to the dataset. The reviews included in this analysis are of initial manuscript submissions. No revised manuscript submissions were included. Article types include research submissions, review articles, brief communications, and view articles. After examination for adequate connectivity through the data, connectivity problems were found with 10 reviews of manuscripts. These were removed from the dataset, reducing the number of reviews to 635 reviews of 301 manuscripts.

## Instrument and Variables

The dataset contains five items that measure aspects of a single construct, publishability. These items were designed to capture important features that should be present in scientific manuscripts desirable for publication and are taken into account during the decision making process. These five criteria are "Novelty," "Clinical Impact," "Scientific Impact," "Definitive," and "Interesting to Specialty." Each reviewer provides ratings on each criterion using a scale of one to five to denote the extent to which the manuscript possesses that criterion. For Novelty, Definitive, and Interesting to Specialty, a rating of one corresponds to "not at all," while a rating of three represents "average," and a rating of five indicates "completely." For Clinical Impact and Scientific Impact, a rating of one, three, and five corresponds to "none," "average," and "immense," respectively. Ratings of two and four are expected to fall between the defined categories. The authors also provide a recommendation of "accept," "major revision," "minor revision," or "reject". The final decision about the manuscript as determined by the editors also is included in the dataset.

## Assumptions about Data and Constructs

The data used in a Generalizability Theory analysis should be ordinal or interval data (Shavelson & Webb, 1991). The observed score is assumed to be made up of the universe score plus sources of error that are assumed to be independent of the universe score. Additionally, all observed behavior, items, and other conditions are assumed to be random samples from the population, if the variable is to be considered random, and not fixed, in the model. For example, reviewers are not randomly sampled, but they are

assumed to be exchangeable for other reviewers. Lastly, error distributions are assumed to be fixed.

Many-Fact Rasch Measurement does not require as many assumptions. One of the main assumptions is that of connectivity through the data. For example, there should be overlap of reviewers reviewing some of the same manuscripts as other reviewers. The model must be unidimensional, meaning that a single latent trait is being measured (e.g., publishability) (Eckes, 2009). Local independence also must hold for Many-Facet Rasch Measurement analyses. For example, a rating given on one item should not affect the rating given on another item of the manuscript rating scale after accounting for the effects of publishability. If local independence does not hold, estimates from the model can be biased, leading to misinterpretation of the data and incorrect decisions based on results.

## Generalizability Theory Analysis

Generalizability Theory analysis was conducted with a two-facet, partially nested design. Manuscript was the object of measurement. Items were one facet. Reviewers were another facet and were nested within manuscript, meaning that each manuscript was rated by different reviewers. The sources of variation in this design were manuscript (m), items (i), and the reviewers nested within manuscript (r:m). Variance components were the manuscript variance component ($\sigma_m^2$), item variance component ($\sigma_i^2$), reviewer nested within manuscript variance component ($\sigma_{r,rm}^2$), the variance component for the manuscript-by-item interaction $\sigma_{mi}^2$), and the variance component for the three-way interaction between manuscript, item, and reviewer plus remaining sources of systematic and unsystematic variation not measured in the study ($\sigma_{ir,mir,e}^2$). Although reviewers

were not randomly sampled, they were assumed to be exchangeable for other reviewers and were considered a random facet. Because this is a nested design, there was no separate variance component for the reviewer facet and no reviewer-by-item interaction (Equation 1). After results were obtained, a Decision Study was conducted to determine the number of items and reviewers needed to improve the generalizability coefficient and the index of dependability (Equations 3 & 5). An additional Decision Study was conducted with items as a fixed facet. urGENOVA software was used for the Generalizability Theory study, and GENOVA software was used for the Decision Study (Brennan, 2001).

### Many-Facet Rasch Measurement Analysis

Many-Facet Rasch Measurement analysis was conducted using the *Facets* program (Linacre, 2014). The rating scale model (Andrich, 1978) was used because the five items were designed to have the same scale structure. Additionally, expert opinion recommends using this model over the partial credit model unless a strong rationale exists for using the partial credit model (Linacre, 2000; Wright, 1998). Because such a rationale has not been developed, the rating scale model was deemed the best choice for this analysis.

Observed averages for manuscript scores, reviewer ratings, and items scores were computed. Reviewer severity measures were obtained to examine reviewer behavior. Manuscript publishability measures corrected for reviewer severity also were produced. Model fit was evaluated for reviewers, manuscripts, and items. Both infit and outfit were assessed, and values outside the range of 0.50 to 1.50 (Linacre, 2002) were flagged for

review and summarized. Item discrimination, the degree to which the score on an item reflects the score on the overall scale, was computed, and negative values and values outside the range of 0.50 to 1.50 were flagged for review and summarized (Linacre, 2014). Item thresholds, representing the points on the theta scale where the likelihood of one rating level is equal to the likelihood of the next rating level, were computed. These thresholds were examined to determine if there was appropriate ordering of difficulties and if there was adequate separation between rating categories. The distribution of difficulties over the theta scale also was noted with the intention of flagging those that were out of order or fell out of the -2.0 to 2.0 range (Myford & Wolfe, 2003). Additional metrics included fair scores, the manuscript scores that would have come from a reviewer with average severity on an item of average difficulty, reviewer fair averages, each reviewer's mean rating adjusted for manuscript publishability and item difficulty, and item fair averages, the average score on each item after adjustment for manuscript publishability and reviewer severity. The reviewer facet was dropped from an additional analysis to evaluate its contribution to the detection of differences in manuscripts.

## Research Question Analyses

### Generalizability Theory

Research Question 1: What proportion of variance in observed scores is attributable to reviewer variation, and how does this compare to the proportion of variance attributable to other sources?

To examine the proportion of variance attributable to reviewer variation and other sources, results from the Generalizability Theory analysis were used. Specifically, a table of variance components of main effects and interactions was produced. Because the study design is partially nested, a variance component for reviewer alone was not available. The variance component associated with reviewer nested within manuscript ($\sigma^2_{r,rm}$) was evaluated instead. The variance component for the object of measurement, manuscript ($\sigma^2_m$), also was evaluated. A variance component for the item facet was available ($\sigma^2_i$), but a variance component for the item-by-reviewer interaction was not available in this study design. The manuscript-by-item interaction variance component ($\sigma^2_{mi}$) was available and was evaluated. The final variance component was the item-by-reviewer interaction confounded with the three-way manuscript-by-item-by-reviewer interaction and other sources of error ($\sigma^2_{ir,mir,e}$).

Research Question 2: Do the results of a Generalizability Theory Decision Study suggest that the conditions of measurement (i.e., number of reviewers and number of items) for manuscript reviews be changed?

To determine whether the results of a Generalizability Theory Decision Study suggest that the conditions of measurement (i.e., number of reviewers and number of items) for manuscript reviews be changed, variance components, generalizability coefficients, and indexes of dependability were calculated for increasing numbers of reviewers and items. The number of reviewers and items required to reach a

generalizability coefficient and an index of dependability of 0.80 were evaluated and interpreted for real-world plausibility. An additional Decision Study was conducted with the item facet fixed, and these results were evaluated.

**Many-Facet Rasch Measurement**

Research Question 3: Do raw publishability scores versus theta scores predict meaningfully different manuscript decision classifications?

To determine whether raw publishability scores versus theta scores result in meaningfully different manuscript decision classifications, results from the Many-Facet Rasch Measurement analysis were used, and theta scores, representing publishability, were obtained for each manuscript. Outside of the Many-Facet Rasch Measurement analysis, average raw total publishability scores were computed for each manuscript by summing the ratings on the five publishability items and taking the average total score across all reviews of that manuscript. Two multinomial logistic regressions were then conducted with the first using average raw total score as the predictor variable and the second using the publishability (theta) measure as the predictor variable. The manuscript decision categories (i.e., accept/minor revision, major revision, and reject) served as the outcome variable. The reject category was used as the reference category. The results of the two models were then interpreted and compared. Results of interest were the odds ratios and 95% confidence intervals for each model and the percentage of correct decision category predictions for each model. Additionally, average raw total scores were plotted against publishability measures to visually depict the effects of adjustment

for rater severity. A polyserial correlation between average raw total score and manuscript decision was explored as a validity coefficient.

Research Question 4: How closely do ranks of the severity measure from each reviewer in a Many-Facet Rasch Measurement analysis compare to ranks of reviewers using average raw ratings from each reviewer?

To examine how closely the ranks of the severity measure of each reviewer in a Many-Facet Rasch Measurement analysis compare to ranks using average raw ratings from each reviewer, a Spearman's rank order correlation and a Pearson correlation were used to compare the reviewer severity measures from the Many-Facet Rasch Measurement analysis and the reviewers' average raw ratings. The proportion of shared variance was computed from these correlations. Additionally, average raw ratings were plotted against reviewer severity measures from the Many-Facet Rasch Measurement analysis to visually depict the association between the two.

# CHAPTER IV

# RESULTS

The results of the analysis of manuscript reviewer ratings are described herein. The section begins with a description of the characteristics of the data. Results of the Generalizability Theory analysis and the Many-Facet Rasch Measurement analysis follow. Next, the applicable results are applied to the research questions, and the findings of these analyses are described.

## Characteristics of the Data

In the data from reviewers who contributed two or more reviews, reviewers used the full range of categories of the five publishability items (Table 1). Each item had responses ranging from the possible minimum of one to the possible maximum of five. The mean rating for each item was around three. The Novelty and Interesting to Specialty items had slightly higher means, suggesting that reviewers more often rate manuscripts higher on these items. The Scientific Impact and Definitive items had the lowest means, suggesting that reviewers may typically give lower ratings on these items. Ratings of one and five were given least often on all of the items except Interesting to Specialty, which received more ratings of five than the other items. A rating of three was the most common rating for all items.

Table 1

Characteristics of Manuscript Rating Items

| Item | Mean (SD) | 1 n (%) | 2 n (%) | 3 n (%) | 4 n (%) | 5 n (%) |
|---|---|---|---|---|---|---|
| Clinical Impact | 2.92 (0.95) | 34 (5.35) | 187 (29.45) | 229 (36.06) | 164 (25.83) | 21 (3.31) |
| Definitive | 2.68 (0.92) | 65 (10.23) | 200 (31.50) | 252 (39.69) | 110 (17.32) | 8 (1.26) |
| Interesting to Specialty | 3.57 (0.90) | 9 (1.42) | 51 (8.03) | 248 (39.05) | 226 (35.59) | 101 (15.91) |
| Novelty | 3.28 (0.90) | 19 (2.99) | 93 (14.65) | 257 (40.47) | 221 (34.80) | 45 (7.09) |
| Scientific Impact | 2.69 (0.88) | 47 (7.40) | 226 (35.59) | 242 (38.11) | 114 (17.95) | 6 (0.95) |

*Note.* N=635 reviews for all items. SD = standard deviation.

Examination of manuscript decisions revealed that some decision categories were used more often than others (Table 2). Manuscripts received editor's decisions of major revision most often and were rejected second most often. Reviewers recommended major revision most often and minor revision second most often. Reviewers recommended acceptance more often than editors but recommended rejection less often than editors. Acceptance decisions were given by editors a very small percentage of the time, and a decision of minor revision was the second least often result. This suggested that some categories should be collapsed in the analysis. Therefore, for the purposes of this study, the accept and minor revision categories were analyzed as one combined category.

Table 2

Manuscript Decisions

| | Reviewer Recommendation N = 635 | Editor Final Decision N = 301 |
|---|---|---|
| Decision Category | Frequency (%) | Frequency (%) |
| Accept | 51 (8.03) | 3 (1.00) |
| Minor Revision | 210 (33.07) | 62 (20.60) |
| Major Revision | 265 (41.73) | 142 (47.17) |
| Reject | 109 (17.17) | 94 (31.23) |

*Note.* Accept and minor revision categories were collapsed in the analyses.

Manuscript total scores on all five items ranged from six to 25 (Figure 1). The average total score was 14.98 (SD = 3.02). Manuscripts received a wide variety of total scores with many of those scores near the mean.

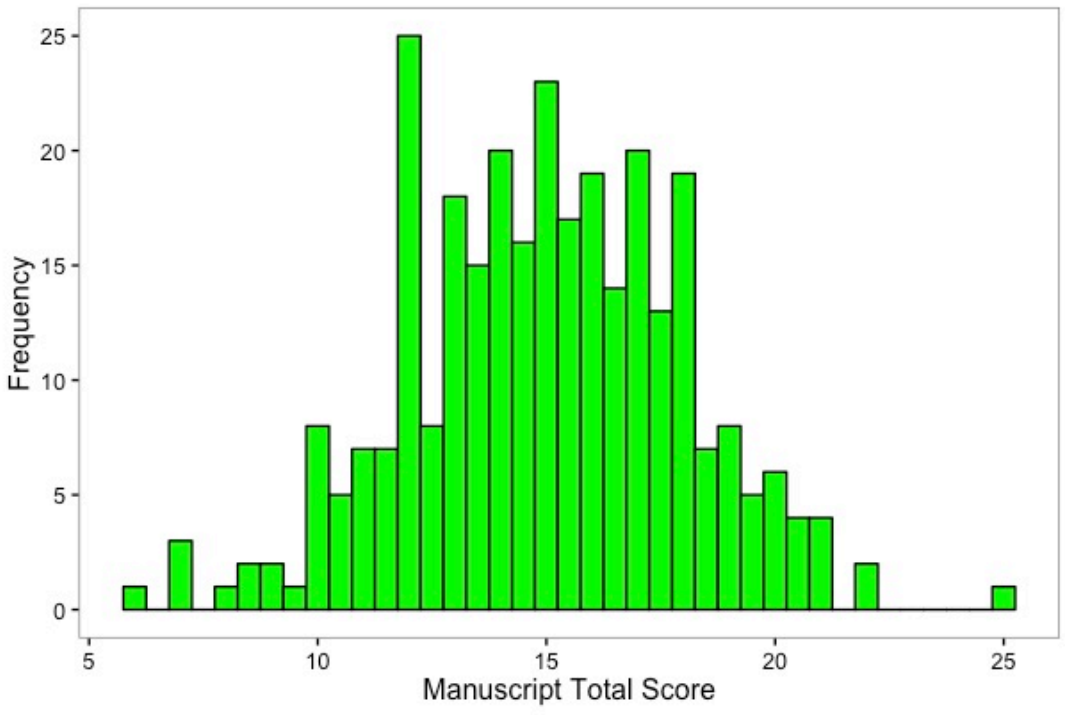


Figure 1. Total Scores for Manuscripts.

## Generalizability Theory Analysis

The manuscript rating data was used to conduct a Generalizability Study. Then, the results of the Generalizability Study were used to conduct a series of Decision Studies. These results are fully described below and later applied to the relevant research questions.

### Generalizability Study

The Generalizability Study produced variance components for each source of variance in the data (Table 3). The manuscript variance component is the second smallest and is equal to 0.1195. This indicates that manuscripts account for 12.21% of the total variance in publishability scores. Reviewers nested within manuscript account for 35.48% of the variance and have the largest variance component of any of the sources of variability. Items account for 15.22% of the total variance, and the manuscript-by-item interaction accounts for 3.54%. The variance component representing the three-way interaction of manuscripts, reviewers, and items plus other sources of error is the second largest, accounting for 33.55% of the total variance.

Table 3

Variance Components

| Source of Variation | df | Mean Squares | Variance Component | Estimated Variance Component | Percentage of Total Variance |
|---|---|---|---|---|---|
| Manuscripts (*m*) | 300 | 3.3977 | $\sigma_m^2$ | 0.1195 | 12.21 |
| Reviewers (*r:m*) | 334 | 2.0648 | $\sigma_{r,rm}^2$ | 0.3473 | 35.48 |
| Items (*i*) | 4 | 94.9769 | $\sigma_i^2$ | 0.1489 | 15.22 |
| Manuscript x Item Interaction *(mi)* | 1200 | 0.4014 | $\sigma_{mi}^2$ | 0.0346 | 3.54 |
| Three-way Interaction & Other Error *(ir,mir,e)* | 1336 | 0.3284 | $\sigma_{ir,mir,e}^2$ | 0.3284 | 33.55 |

**Decision Studies**

The variance components from the Generalizability Study were used to estimate variance components for hypothetical reviewers and item numbers and to estimate generalizability and dependability coefficients for the existing median number of manuscripts and number of items. Results for small increases in both reviewers and items are displayed in Table 4. As the number of both reviewers and items increased, the variance components for reviewers nested within manuscript, items, manuscript-by-item interaction, and residual variance decreased and continued to decrease as reviewers and items increased. The variance component associated with manuscripts was the largest variance component after each of these changes. The variance component for reviewer nested within manuscript and the residual variance component decreased when the number of reviewers increased while items remained the same. This reflects the large reviewer nested within manuscript effect seen in the Generalizability Study and suggests

that changing the number of reviewers has more impact than changing the number of items.

The generalizability coefficient for the current number of reviewers and items was 0.3590 (Table 4). This low number suggests that the scores of reviewers cannot be considered reliable for making relative decision about manuscript publishability (i.e., if some manuscripts are more publishable compared to others). The index of dependability (phi coefficient) for the current number of items and reviewers also was low at 0.3295, implying that reviewer manuscript ratings cannot be considered reliable for making absolute decisions about manuscript publishability (i.e., whether manuscripts meet some level of publishability, regardless of the status of other manuscripts).

Table 4

Decision Study

| | G Study | Alternative Decision Studies | | | | | |
|---|---|---|---|---|---|---|---|
| Source of | $n'_r =$ 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Variation | $n'_i =$ 5 | 6 | 6 | 8 | 8 | 10 | 10 |
| $\sigma_m^2$ | 0.1195 | 0.1195 | 0.1195 | 0.1195 | 0.1195 | 0.1195 | 0.1195 |
| $\sigma_{r,rm}^2$ | 0.3473 | 0.1158 | 0.0868 | 0.0695 | 0.0579 | 0.0496 | 0.0434 |
| $\sigma_i^2$ | 0.1489 | 0.0248 | 0.0248 | 0.0186 | 0.0186 | 0.0149 | 0.0149 |
| $\sigma_{mi}^2$ | 0.0346 | 0.0058 | 0.0058 | 0.0043 | 0.0043 | 0.0035 | 0.0035 |
| $\sigma_m^2$ | 0.3284 | 0.0183 | 0.0137 | 0.0082 | 0.0068 | 0.0047 | 0.0041 |
| $\sigma_{Rel}^2$ | 0.2134 | 0.1398 | 0.1063 | 0.0820 | 0.0691 | 0.0578 | 0.0510 |
| $\sigma_{Abs}^2$ | 0.2432 | 0.1646 | 0.1311 | 0.1006 | 0.0877 | 0.0727 | 0.0659 |
| $\rho^2$ | 0.3590 | 0.4610 | 0.5293 | 0.5931 | 0.6338 | 0.6742 | 0.7010 |
| $\Phi$ | 0.3295 | 0.4207 | 0.4769 | 0.5430 | 0.5769 | 0.6219 | 0.6447 |

*Note*. $\boldsymbol{\rho^2}$ represents the generalizability coefficient; $\boldsymbol{\Phi}$ represents the index of dependability, which is also called the phi coefficient.

Both the generalizability coefficient and the index of dependability increased in the alternative Decision Studies, suggesting that increasing the number of reviewers and items will improve the reliability with which decisions about whether to publish manuscripts can be made (Table 4). The generalizability coefficient increased at a faster rate than the index of dependability, but neither reached an acceptable level with the prespecified numbers of reviewers and items in the Decision Studies. For this reason, additional Decision Studies were undertaken to explore the number of reviewers and items that would be necessary to achieve appropriate reliability for decision making (Appendix A). Because the number of items did not appear to have as much effect as the number of reviewers, items were not increased and the numbers of five to ten items were used in the additional analyses.

Figure 2 shows how an increase in number of reviewers affects the generalizability coefficient for different numbers of items. For the generalizability coefficient to reach the acceptable level of 0.80, a minimum of seven items and 16 reviewers is required. At a low number of reviewers, different numbers of items do not appear to influence the generalizability coefficient as much as at higher numbers of reviewers. Additionally, increases in reviewers lead to greater changes in the generalizability coefficient when the number of reviewers is below 10. After this point, the effects on the generalizability coefficient become subtler.

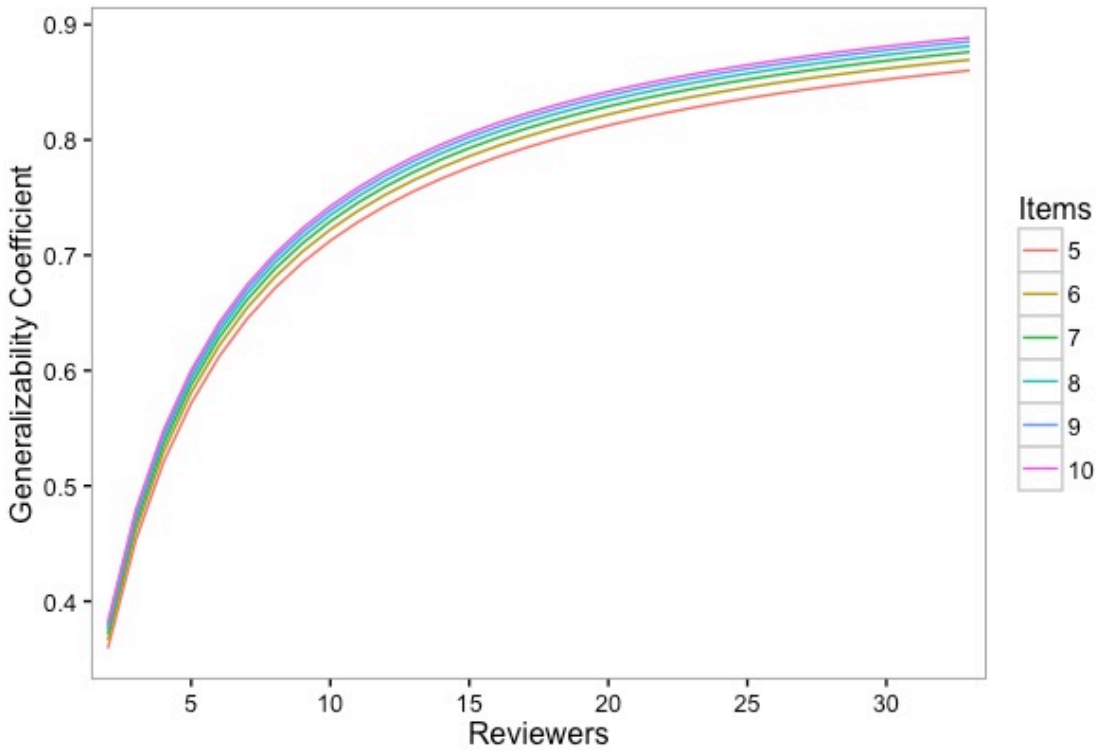


Figure 2. Generalizability Coefficient at Increasing Numbers of Reviewers and Items.

Figure 3 demonstrates the effects of an increase in number of reviewers on the index of dependability for different numbers of items. For the index of dependability to reach the acceptable level of 0.80, a minimum of 10 items and 33 reviewers is required. Like the generalizability coefficient, at low number of reviewers, different numbers of items do not appear to influence the generalizability coefficient as much as at higher numbers of reviewers. Increases in reviewers lead to greater changes in the generalizability coefficient when the number of reviewers is below seven or eight. After this point, the effects on the generalizability coefficient increase at a slower rate.

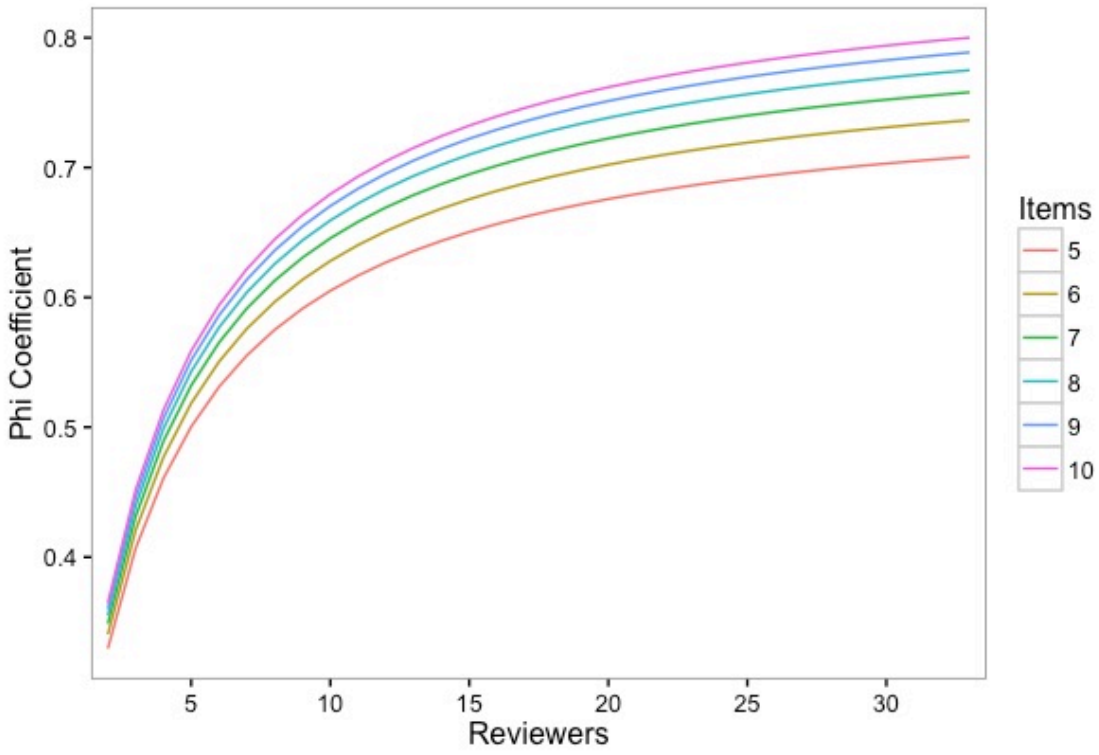


Figure 3. Index of Dependability (Phi Coefficient) at Increasing Numbers of Reviewers and Items.

The number of items seems to have a greater effect on the index of dependability than on the generalizability coefficient. This is due to the presence of the item main effect variance component in the calculation of the index of dependability but not in the generalizability coefficient, which includes only interaction variance components.

**Items as a Fixed Facet**

An additional attempt at understanding and improving reliability of reviewer scores was made by fixing the item facet and computing Decision Studies with this design. By fixing the item facet, the five included items are considered to be the only items that will be used to assess publishability in the universe of generalization. With the

item facet fixed at five, the generalizability coefficient for a study with two reviewers was 0.4077, and the index of dependability was 0.4077. Both are the same with the item facet fixed because fixing item variance removes this effect from the calculation of absolute error variance and, therefore, the index of dependability. These equations become the same as the equations for relative error variance and the generalizability coefficient. When both equations are calculated in the same manner, the generalizability coefficient and index of dependability will be the same. From these results, for both coefficients to reach 0.80, 12 reviewers would be required (Figure 4).

Fixing the item facet increased the generalizability coefficient and the index of dependability for the existing number of reviewers (n = 2). Under scenarios of increasing numbers of reviewers, the two coefficients reached the level of 0.80 with fewer reviewers than when the item facet was random (Appendix B). Removing item variability appears to improve the reliability of scores obtained from manuscript reviewers.

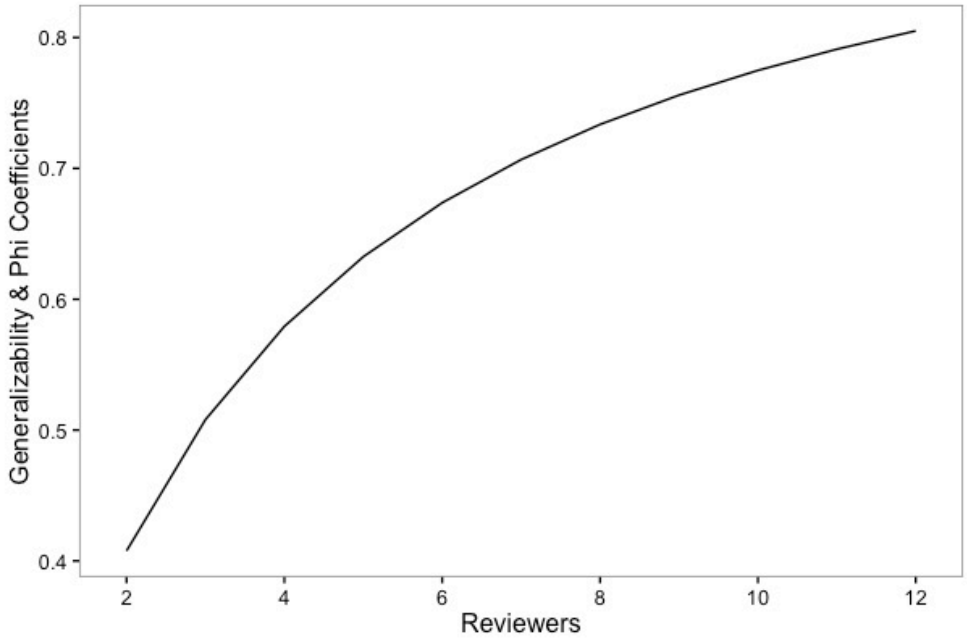


Figure 4. Generalizability Coefficient and Index of Dependability (Phi Coefficient) at Increasing Numbers of Reviewers and Items. Both coefficients generate the same line in this analysis.

## Many-Facet Rasch Measurement Analysis

The reviews' scores on the five items were used to produce results from the Many-Facet Rasch Measurement analysis. The analysis provided results for manuscripts, reviewers, and items. Model fit information also was obtained. The findings of this analysis are described here and later applied to the appropriate research questions.

### Manuscript Facet

The observed average score on the publishability items was 3.00 (SD = 0.60) (Table 5). This means that manuscripts received an average rating of three across all items. The fair average adjusted for reviewer severity and item difficulty was slightly higher, which suggests that manuscripts may be rated slightly higher if ratings were based

on the average reviewer and the average item difficulty. The average publishability (theta) measure was 0.14 (SD = 1.66) with a wide range of publishability levels from -4.88 to 8.77. This indicates much variety in the quality of the reviewed manuscripts.

Table 5

Manuscript Many-Facet Rasch Measurement Analysis

| | Mean (SD) | Range |
|---|---|---|
| Observed Average | 3.00 (0.60) | 1.20 to 5.00 |
| Fair Average | 3.07 (0.65) | 1.23 to 4.99 |
| Publishability Measure | 0.14 (1.66) | -4.88 to 8.77 |
| Standard Error | 0.53 (0.16) | 0.32 to 1.89 |
| Infit (MS) | 0.90 (0.64) | 0.09 to 3.89 |
| Outfit (MS) | 0.91 (0.64) | 0.09 to 3.78 |
| Discrimination | 1.10 (0.68) | -1.99 to 1.96 |
| Separation ratio | 2.81 | |
| Reliability of separation | 0.89 | |

*Note.* SD = standard deviation; MS = mean-square.

Fit statistics suggested that there were inconsistencies in the fit of the model for the included manuscripts. Infit statistics revealed that there were many unexpected scores for manuscripts. Average mean-square infit was 0.90 (SD = 0.64), and infit statistics for individual manuscripts ranged from 0.09 to 3.89 (Figure 5). Using a range of 0.5 to 1.5 for acceptable infit (Linacre, 2002), 136 (45.18%) manuscripts exhibited infit problems. Of these, 91 (30.23%) fell below 0.5, indicating too little variation in scores, and 45 (14.95%) fell above 1.5, indicating excess unmodeled variation (Linacre, 2002). Outfit statistics revealed many instances of typically consistent manuscripts receiving unexpected scores. Average mean-square outfit was 0.91 (SD = 0.64), and outfit statistics for individual manuscripts ranged from 0.09 to 3.78 (Figure 6). Using the

0.5 to 1.5 range (Linacre, 2002), 138 (45.85%) of manuscript exhibited outfit problems. Of these, 91 (30.23%) fell below 0.5, suggesting too little variation in scores, and 47 (15.61%) fell above 1.5, suggesting excess unmodeled variation (Linacre, 2002). Discrimination problems also were present in 143 (47.50%) of manuscripts. Of these, 20 (6.64%) manuscripts had negative discrimination values. Another 27 (8.97%) manuscripts had values in the range of 0 to < 0.5, and 96 (31.89%) manuscripts had values above 1.5. These problems suggest that manuscript scores do not strongly distinguish raters from each other or items from each other.

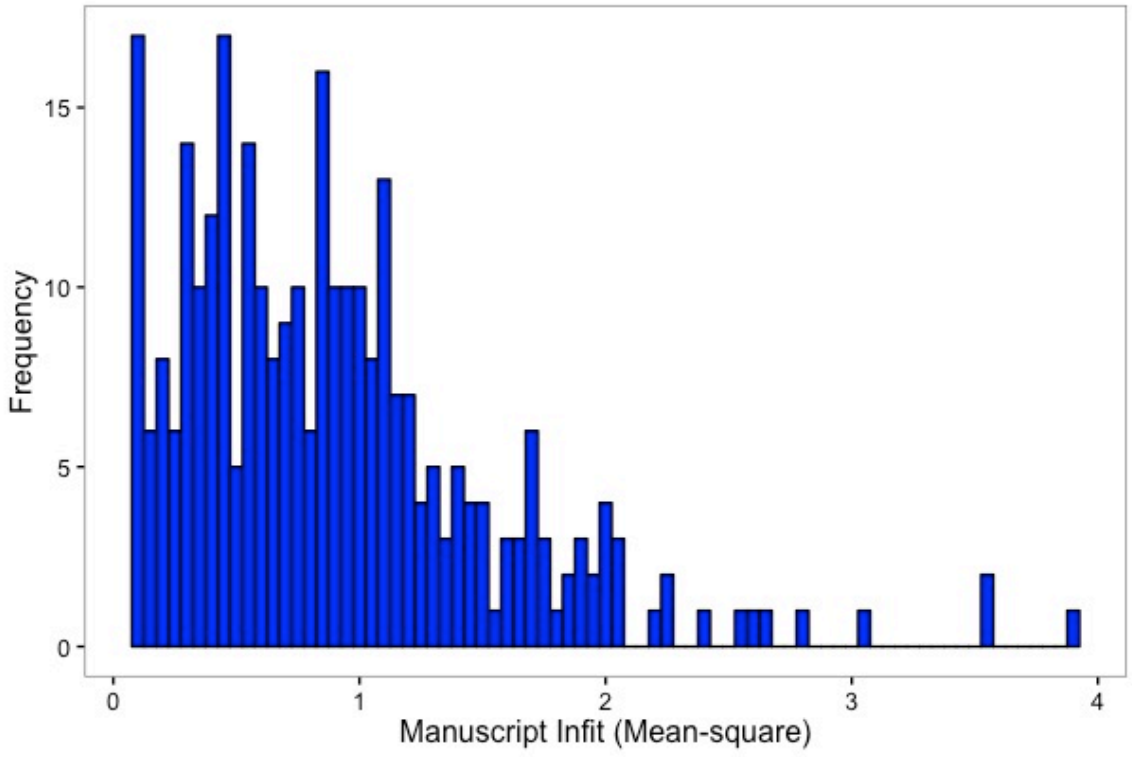


Figure 5. Mean-Square Infit for Each Manuscript.

The separation ratio, a measure of true standard deviation over average measurement error was 2.81 (Table 5). This indicates the number of distinguishable levels of publishability scores that would occur in a normally distributed sample that had the same true standard deviation as this sample (Linacre, 2014). The reliability of separation index of 0.89 indicates that manuscripts are reliably different from each other and approximates true variance over observed variance.

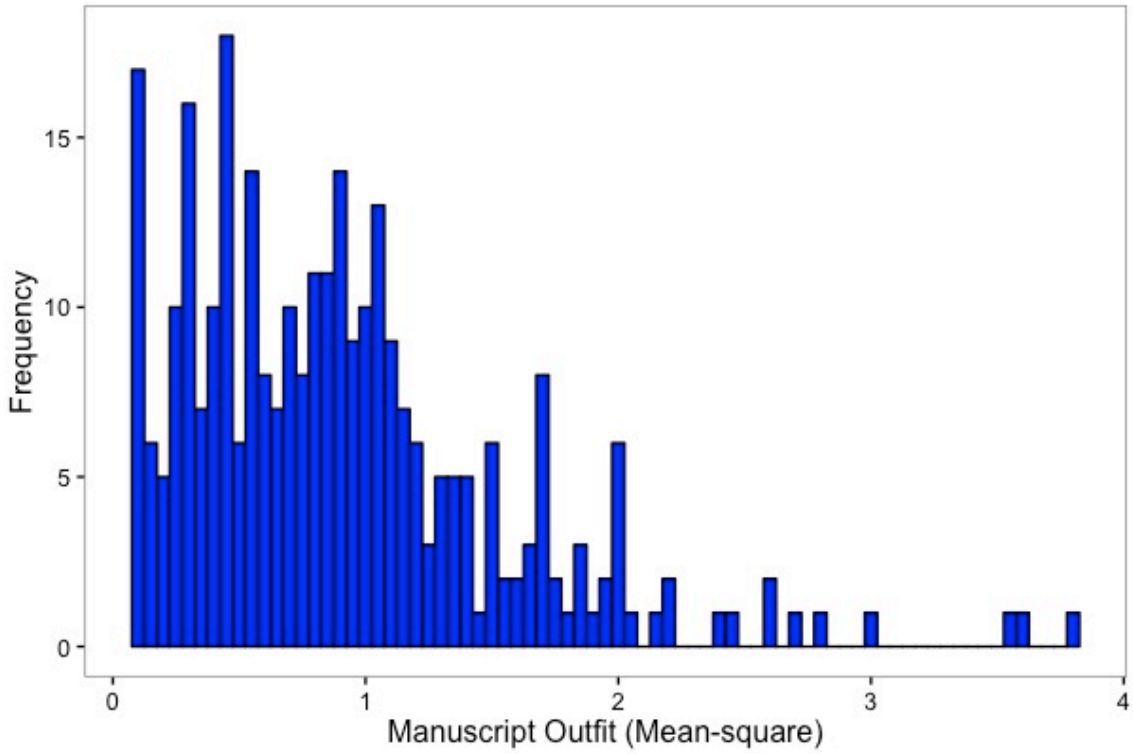


Figure 6. Mean-Square Outfit for Each Manuscript.

**Reviewer Facet**

The observed average rating the reviewers provided was 3.11 (SD = 0.51) (Table 6). The fair average of ratings adjusted for manuscript publishability and item difficulty

was slightly lower, suggesting that reviewers would provide lower ratings if their ratings were based on the average manuscript and the average item difficulty. The average reviewer severity measure was 0.00 (SD = 1.43) with a wide range of severity levels from -5.05 to 3.39. This reveals variety in the severity levels of the included reviewers.

Table 6

Reviewer Many-Facet Rasch Measurement Analysis

| | Mean (SD) | Range |
|---|---|---|
| Observed Average | 3.11 (0.51) | 1.20 to 4.30 |
| Fair Average | 3.06 (0.58) | 1.68 to 4.80 |
| Reviewer Severity Measure | 0.00 (1.43) | -5.05 to 3.39 |
| Standard Error | 0.43 (0.10) | 0.05 to 0.84 |
| Infit (MS) | 1.13 (0.62) | 0.14 to 2.82 |
| Outfit (MS) | 1.14 (0.64) | 0.14 to 3.34 |
| Discrimination | 0.87 (0.65) | -0.87 to 1.93 |
| Separation ratio | 3.07 | |
| Reliability of separation | 0.90 | |

*Note.* SD = standard deviation; MS = mean-square.

Fit statistics suggested that there were some inconsistencies in the fit of the model for the included reviewers. Infit statistics indicated that reviewers provided several unexpected ratings. Average mean-square infit was 1.13 (SD = 0.62), and infit statistics for individual reviewers ranged from 0.14 to 2.82 (Figure 7). Using a range of 0.5 to 1.5 for acceptable infit (Linacre, 2002), 57 (41.30%) reviewers exhibited infit problems. Of these, 21 (15.22%) fell below 0.5, indicating too little variation, and 36 (26.09%) fell above 1.5, indicating excess unmodeled variation (Linacre, 2002). Outfit statistics revealed several instances of typically consistent reviewers providing unexpected scores. Average mean-square outfit was 1.14 (SD = 0.64), and outfit statistics for individual

reviewers ranged from 0.14 to 3.34 (Figure 8). Using the 0.5 to 1.5 range (Linacre, 2002), 56 (40.58%) of reviewers exhibited outfit problems. Of these, 21 (15.22%) fell below 0.5, suggesting too little variation, and 35 (25.36%) fell above 1.5, suggesting excess unmodeled variation (Linacre, 2002).

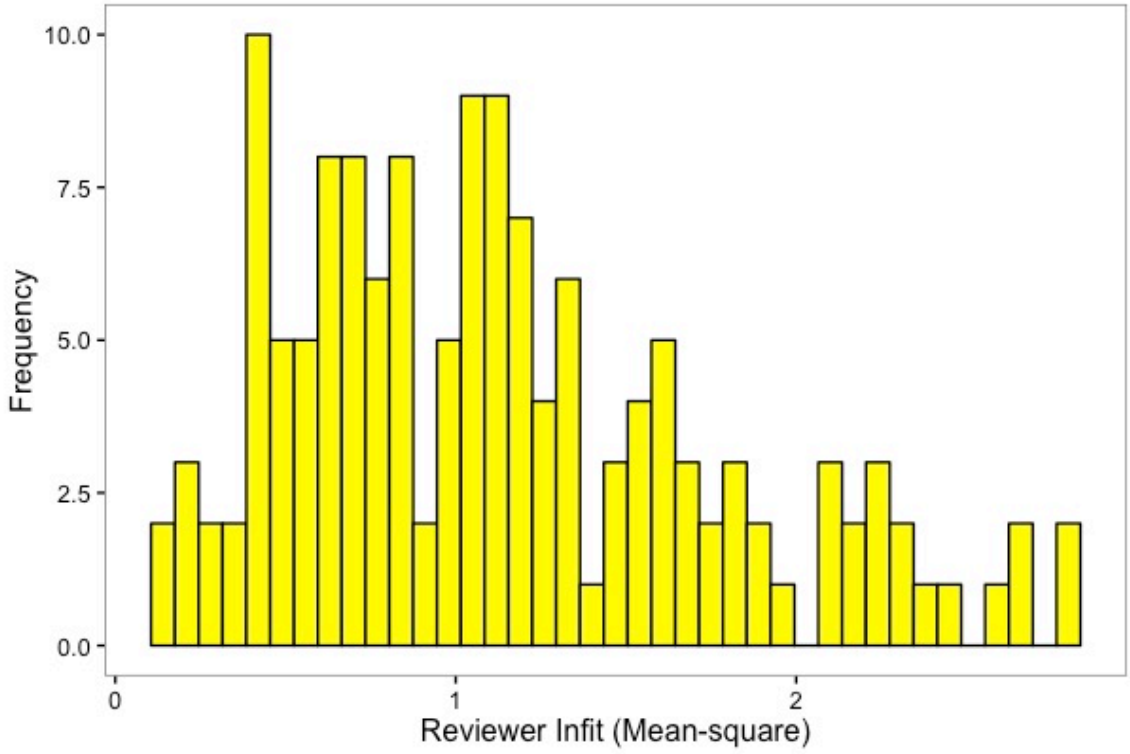


Figure 7. Mean-Square Infit for Each Reviewer.

Discrimination problems were present in 59 (42.75%) of reviewers. Of those, 16 reviewers had negative discrimination values, 20 (14.49%) reviewers had values in the range of 0 to < 0.5, and 23 (16.67%) reviewers had values above 1.5. Problems with reviewer discrimination indicate trouble with the ability of reviewers to discriminate among manuscripts of different levels of quality.

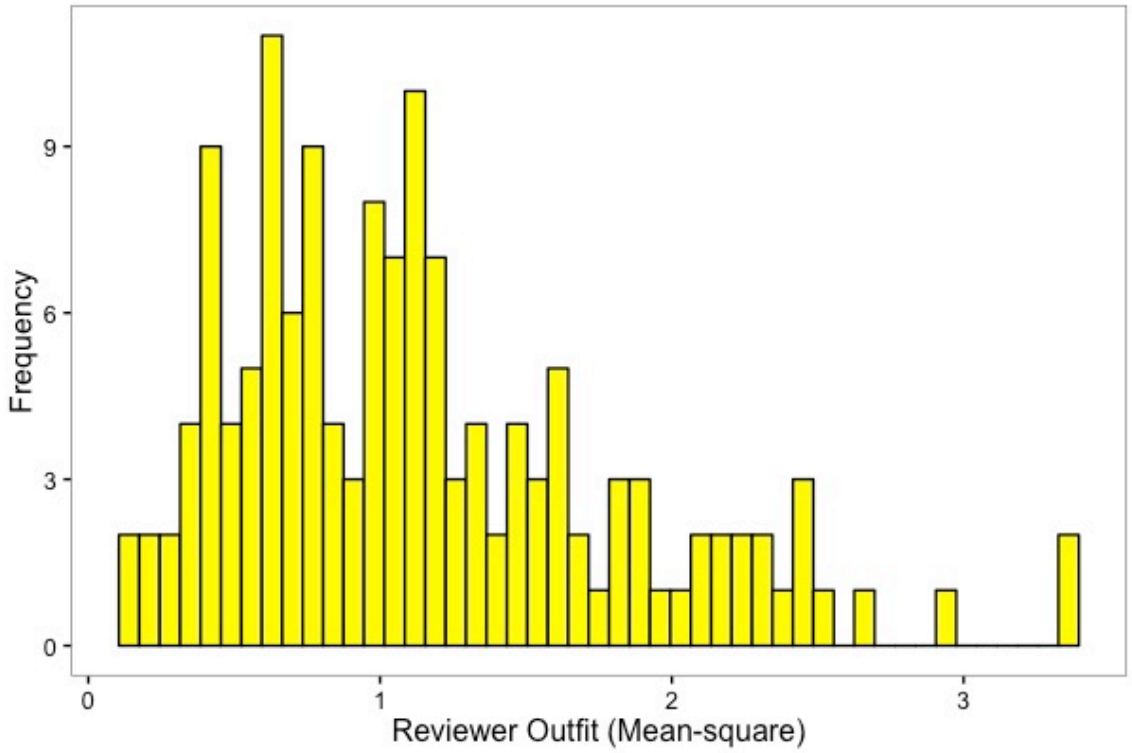


Figure 8. Mean-Square Outfit for Each Reviewer.

The separation ratio, the measure of true standard deviation over average measurement error was 3.07 (Table 6). This indicates the number of distinguishable levels of reviewer severity that would occur in a normally distributed sample that had the same true standard deviation as this sample (Linacre, 2014). The reliability of separation index of 0.90 indicates that reviewers are reliably different from each other. This means that interrater reliability is low, and reviewers do not provide similar ratings for the same manuscripts.

**Item Facet**

For each of the five items, the fair average adjusted for manuscript publishability and reviewer severity slightly increased over the observed average (Table 7). This suggests that item scores may be higher if they were based on the average manuscript and the average reviewer. The Interesting to Specialty item was the least difficult, and the Definitive item was the most difficult. These differences in item difficulty show that obtaining high scores on the items in the publishability scale is more difficult for some items than for others.

Table 7

Item Many-Facet Rasch Measurement Analysis

| | Clinical Impact | Definitive | Interesting to Specialty | Novelty | Scientific Impact |
|---|---|---|---|---|---|
| Observed Average | 2.92 | 2.68 | 3.57 | 3.28 | 2.69 |
| Fair Average | 2.96 | 2.70 | 3.63 | 3.34 | 2.72 |
| Difficulty Measure | 0.27 | 0.87 | -1.35 | -0.62 | 0.83 |
| Standard Error | 0.06 | 0.06 | 0.06 | 0.06 | 0.06 |
| Infit (MS) | 0.97 | 0.98 | 1.02 | 1.18 | 0.80 |
| Outfit (MS) | 1.00 | 0.98 | 1.02 | 1.18 | 0.79 |
| Discrimination | 1.03 | 1.02 | 0.98 | 0.80 | 1.22 |
| Separation ratio | 13.70 | | | | |
| Reliability of separation | 0.99 | | | | |

*Note.* MS = mean-square.

Infit and outfit indices fell within the range of 0.5 to 1.5 for all items (Linacre, 2002). This indicates that all items fit well with the model, and this adequate fit supports the decision to use the rating scale model in the analysis (Andrich, 1978; Linacre, 2000;

Wright, 1998). Discrimination values also were appropriate for all items, suggesting that scores on the items reflect scores on the overall scale.

The separation ratio, the measure of true standard deviation over average measurement error was 13.07 (Table 6). This indicates the number of distinguishable levels of item difficulty that would occur in a normally distributed sample that had the same true standard deviation as this sample (Linacre, 2014). The reliability of separation index of 0.99 indicates that items are reliably different from each other.

Figure 9 displays the overall item characteristic curves for the manuscript publishability scale. Each possible rating from one to five is represented by a curve in the figure, and these curves correctly appear in numbered order. From these curves, it is possible to understand the manuscript publishability levels at which receiving different scores on the items are possible. The probability of moving from a score of one to a score of two increases around a publishability level of negative four. The probability of moving from a score of two to a score of three increases around a publishability level of negative one. Next, the probability of moving from a score of three to a score of four increases around a publishability level of one. Finally, the probability of moving from a score of four to a score of five increases around a publishability level of four.

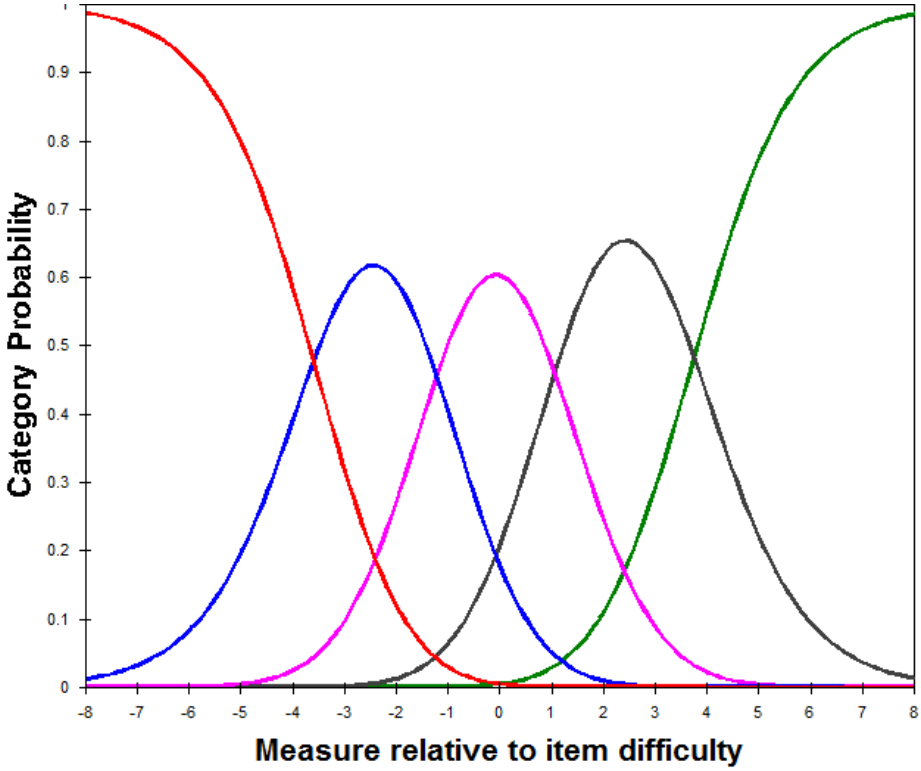


Figure 9. Item Characteristic Curves for Manuscript Publishability Scale.

Figure 10 shows the empirical item characteristic curve constructed from the observations. Some differences from the modeled curves appear in score levels one and two, and it appears that at very low levels of publishability, the probability of receiving either one of these scores is high.

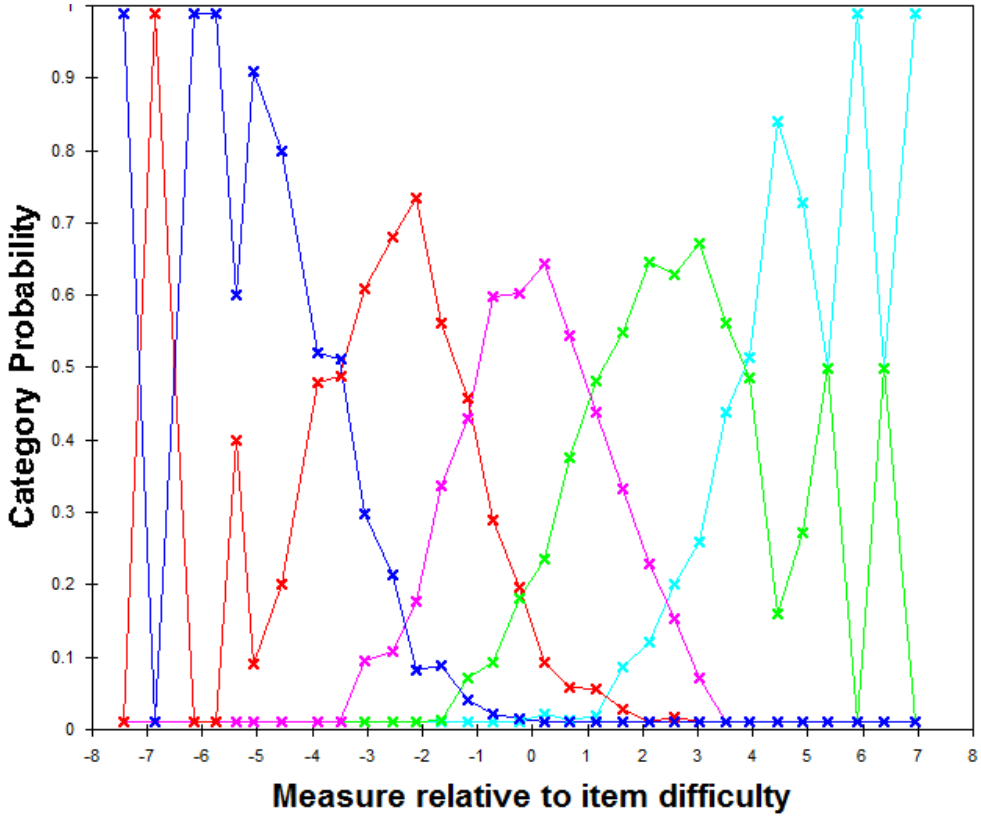


Figure 10. Empirical Category Curves for Manuscript Publishability Scale.

Figure 11 shows the predicted score for manuscripts relative to item difficulty. The dashed lines represent the 0.5 probability thresholds and the publishability levels to which these correspond. These levels are the same as those in Figure 9.

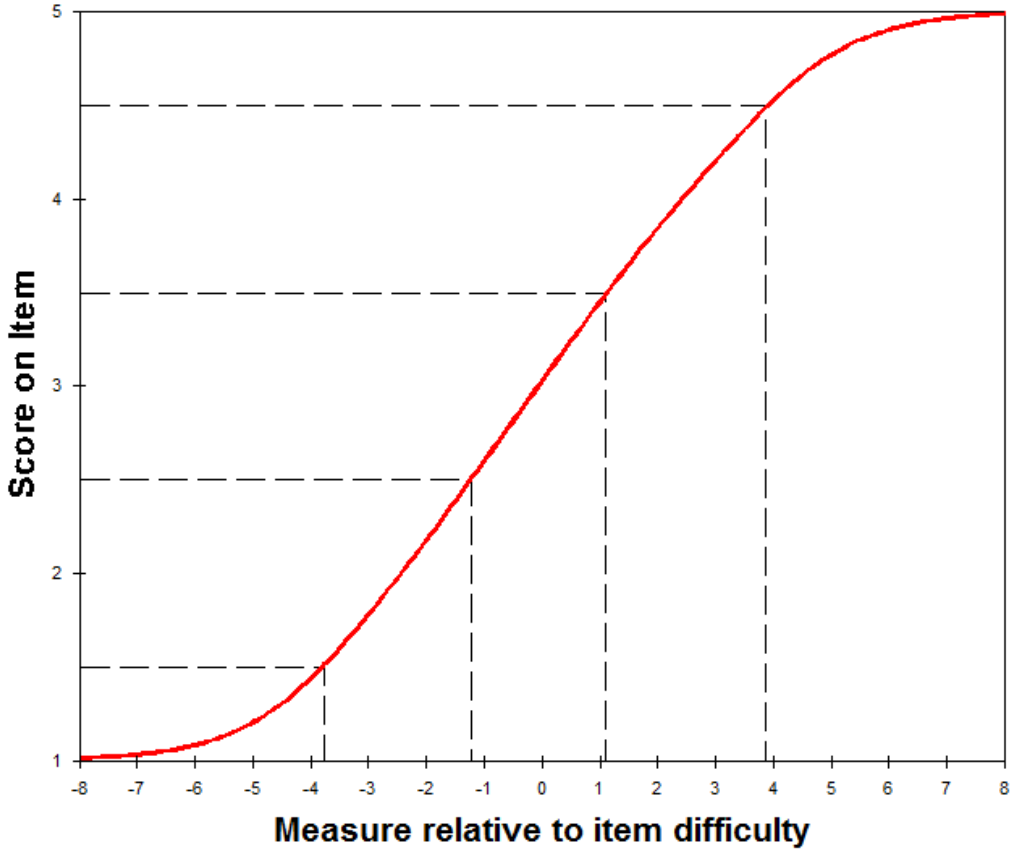


Figure 11. Expected Score Item Characteristic Curve.

Figure 12 shows the empirical item characteristic curve based on the observations. Differences from the modeled curve are apparent in the very low publishability range and the score levels of one and two. Other differences are seen in the very high publishability range and the score levels of four and five. At these extremes, other sources of variance not included in the model may be affecting the observations.

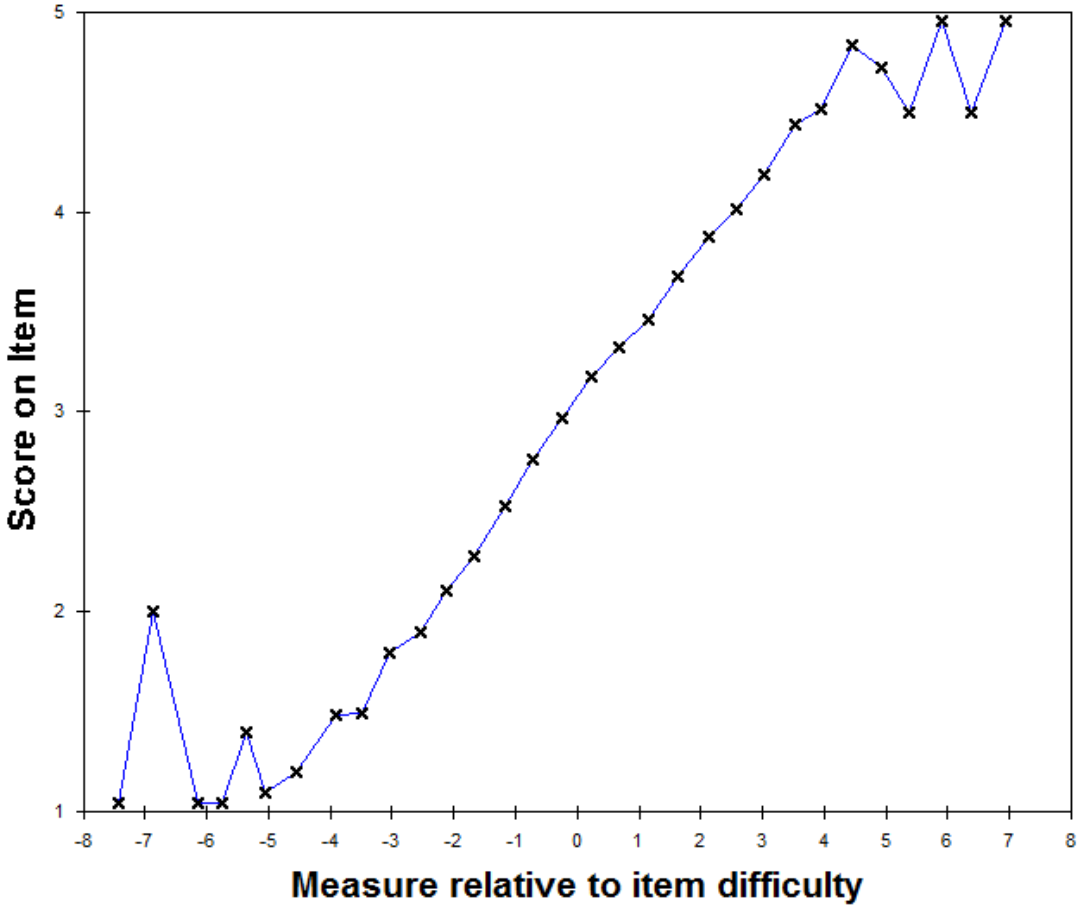


Figure 12. Empirical Item Characteristic Curve.

Figure 13, the conditional probability curve, represents the conditional probabilities of observing adjacent categories. The curves are Rasch dichotomous ogives that cross the 0.5 probability line at the point where the probability of a score in the category is equal to the probability of a score in the next highest category. For example, the line corresponding to a score of one crosses the 0.5 probability line at a publishability level of negative four. This corresponds to thresholds seen in Figure 9.

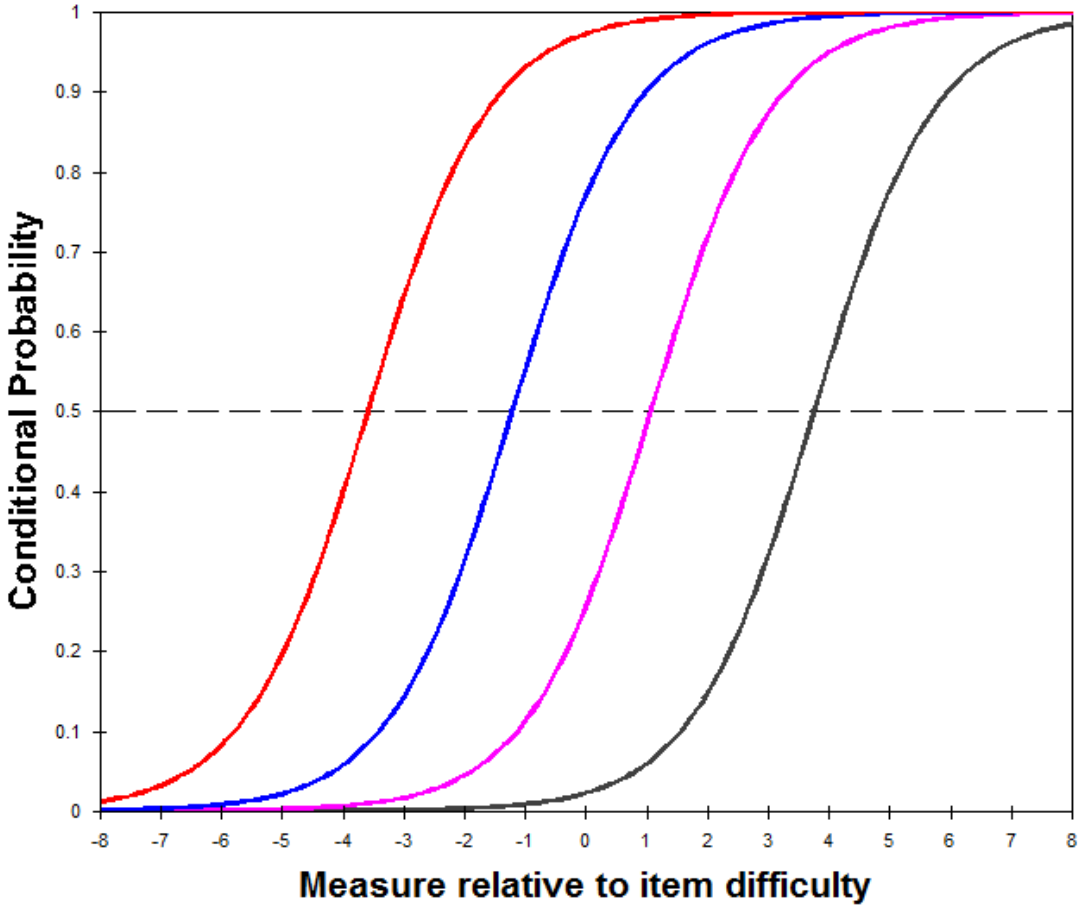


Figure 13. Conditional Probabilities.

In Figure 14, the cumulative probabilities are displayed. Each curve represents the probability of a manuscript with a certain publishability level being observed in that category or the categories below that category. For example, the blue curve (second from the left) represents the probability of being observed in score category two or score category one.

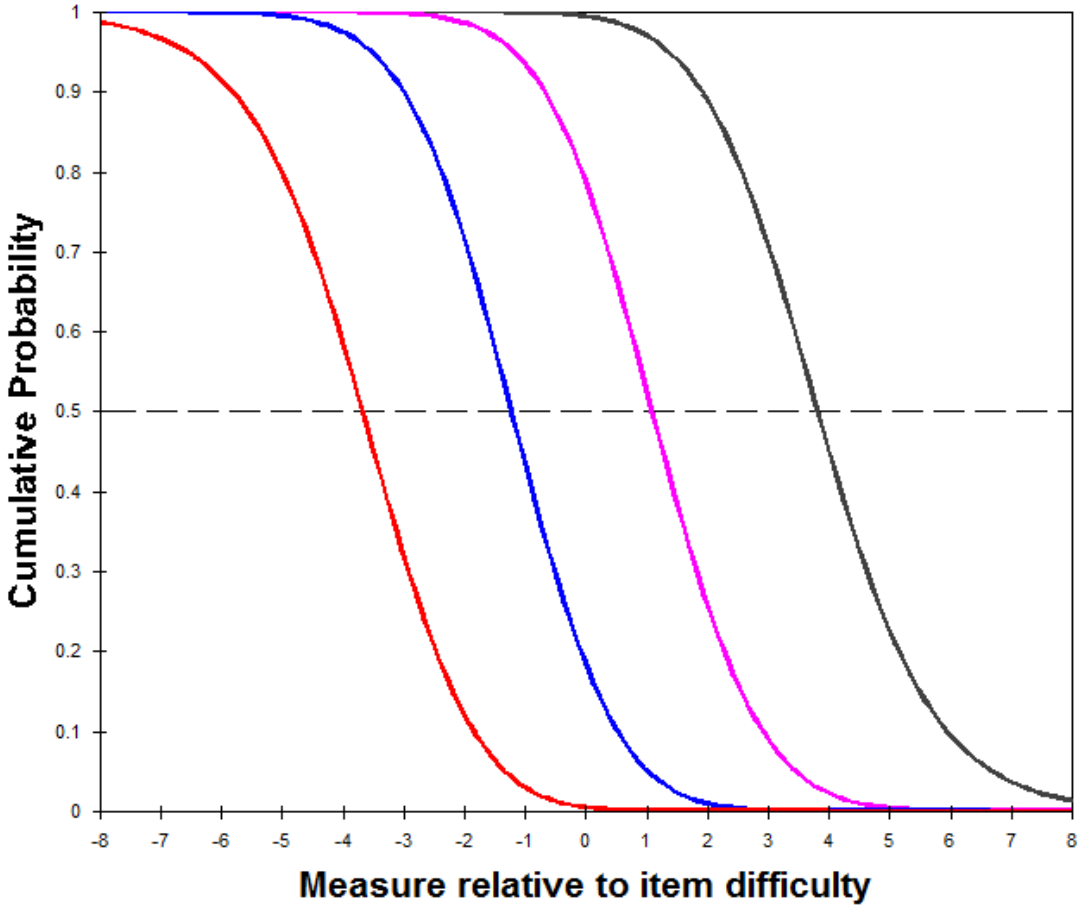


Figure 14. Cumulative Probabilities.

In Figure 15, item information for the range of manuscript publishability levels is shown. The highest level of information occurs around the negative one publishability level. However, information is similarly high from the negative four to positive four publishability range. This indicates that the items provide similar information across the theta range and do not provide higher information at certain levels of publishability.

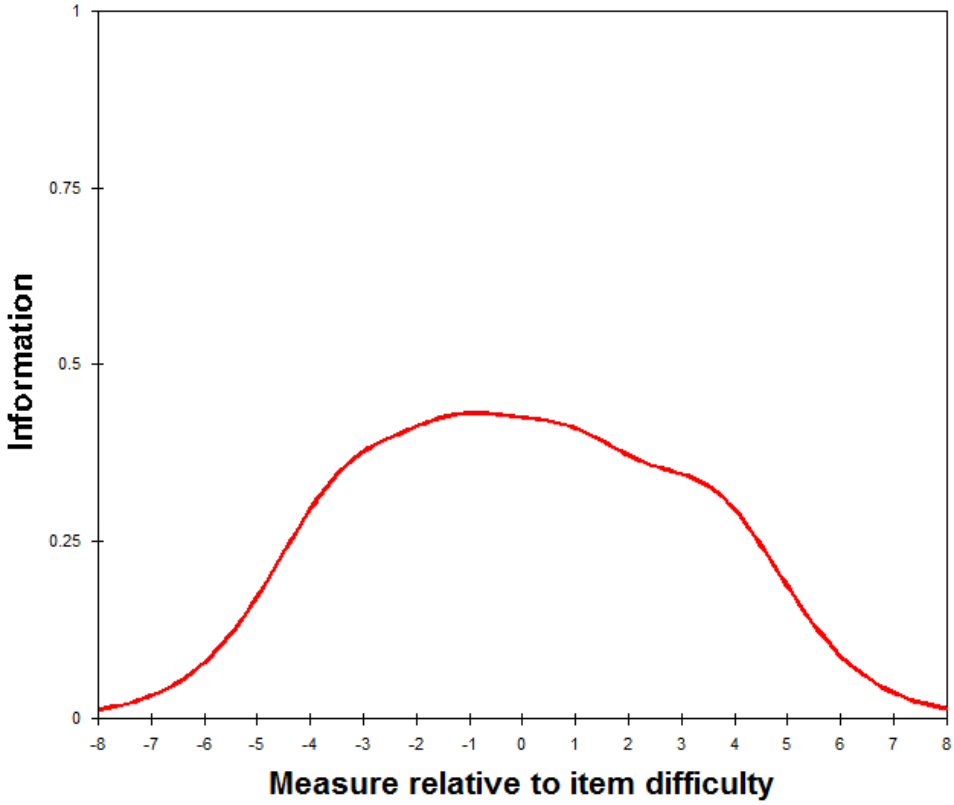


Figure 15. Information Curve for Manuscript Publishability Items.

Likewise, Figure 16, displays information curves for each score category in the publishability scale. Higher levels of information are apparent for categories two, three, and four, but differences between all categories are subtle. This corresponds to Figure 15 and suggests that the categories provide similar information across the theta range and provide only slightly higher information at certain levels of publishability.

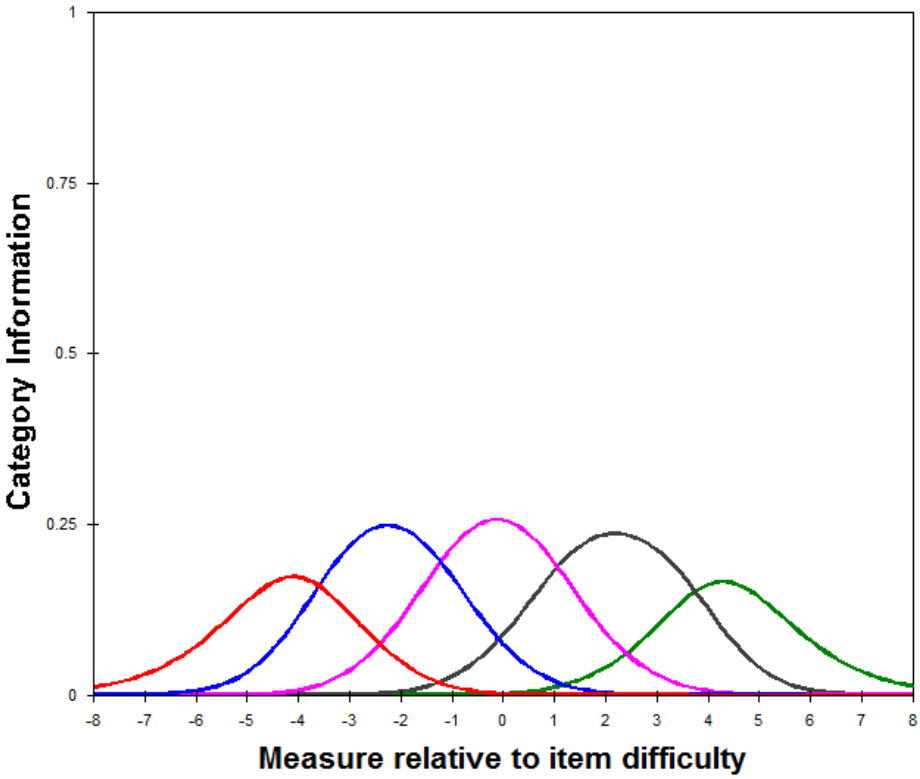


Figure 16. Information Curves for Each Score Category of the Manuscript Publishability Scale.

**Facet Comparisons**

The findings of the analysis of the three facets suggest more variability and less model fit for manuscripts and reviewers compared to items. The manuscript facet exhibited numerous infit and outfit problems and a high standard deviation for the publishability measure (SD = 1.66; Table 5). This suggests that manuscripts were very different from each other and that the model may not have been an appropriate fit for all manuscripts. Similarly, the reviewer facet exhibited fit problems and a high standard deviation for the reviewer severity facet (SD = 1.43; Table 6). Although these issues were slightly less severe than those of the manuscript facet, they still suggest a high level

of variability among reviewers and a poor fit of some reviewers to the model. The item facet was far less problematic and showed no fit problems. The standard deviation of the item difficulty measure was lower than those of the other facets (mean = 0.00, SD = 0.97). The items appeared to fit the model and did not show extreme variability.

### Analysis without Reviewer Facet

To explore the effects of removing the reviewer facet on manuscript reliability of separation, an analysis without the reviewer facet was conducted. The reliability of separation index was 0.89 with the reviewer facet included (Table 5). Without the reviewer facet, reliability dropped to 0.83. This suggests that the reviewer facet adds to the ability to detect the differences in manuscripts and contributes to increasing the ratio of true variance to observed variance.

## Research Question Analyses

To address the four research questions, the results of the Generalizability Theory analysis and the Many-Facet Rasch Measurement analysis are applied in the context of the research questions. Generalizability Theory results are used to address Research Question 1 and Research Question 2. Many-Facet Rasch Measurement results are used to address Research Question 3 and Research Question 4.

### Generalizability Theory

Research Question 1: What proportion of variance in observed scores is attributable to reviewer variation, and how does this compare to the proportion of variance attributable to other sources?

Because reviewers are nested within manuscripts, a separate reviewer variance component cannot be obtained, and the reviewer nested within manuscript variance component must be interpreted. The proportion of variance attributable to reviewer nested within manuscript variation is 35.48% with variance component $\sigma^2_{r,rm}$ = 0.3473 (Table 3). This means that a large proportion of the variance in observed manuscript scores is due to the reviewer nested within manuscript effect, and reviewer behavior differed from one manuscript to another. This effect includes overall differences in reviewer severity as well as the interaction of reviewer severity with the manuscript effect. Consequently, both relative and absolute sources of error variance are involved.

Compared to the other sources of variance, the reviewer nested within manuscript variance is the largest. Manuscripts account for 12.21% of the total variance in observed scores ($\sigma^2_m$ = 0.1195). This suggests that manuscripts differ in their quality. The proportion of variance in observed scores that can be attributed to items is 15.22% ($\sigma^2_i$ = 0.1489), suggesting that items differ in their difficulty. The manuscript-by-item interaction accounts for 3.54% of the total variance ($\sigma^2_{mi}$ = 0.0346) and indicates that the relative standing of manuscripts differed slightly by item. The last remaining source of variance, the three-way interaction confounded with the item-by-rater interaction and other sources or error accounts for 33.55% of the total variance in observed scores ($\sigma^2_{ir,mir,e}$ = 0.3284). Because this variance component is so large, a substantial amount of variability in total scores must be due to these interactions and other sources of unmeasured variability. The nested reviewer effect precludes the item-by-reviewer

interaction. For this reason, whether the relative standing of items differed by reviewers cannot be known.

Research Question 2: Do the results of a Generalizability Theory Decision Study suggest that the conditions of measurement (i.e., number of reviewers and number of items) for manuscript reviews be changed?

The generalizability coefficient ($\rho^2 = 0.3590$; Table 4) and the index of dependability ($\Phi = 0.3259$) from the original study design based on two reviewers and five items are not high enough to be used for decision making and suggest that the conditions of measurement be changed. In the alternative Decision Studies, increasing the number of reviewers and items to realistic amounts improved the coefficients but not enough to reach acceptable levels (Table 4). From these Decision studies, the number of reviewers appeared to have a greater impact on the coefficients than the number of items, so the number of reviewers was increased until the coefficients each reached 0.80.

For the generalizability coefficient to reach the acceptable level of 0.80, a minimum of seven items and 16 reviewers was required. For the index of dependability to reach the acceptable level of 0.80, a minimum of 10 items and 33 reviewers was required. When the item facet was fixed at five, the generalizability coefficient and index of dependability for a study with two reviewers were both 0.4077. For both coefficients to reach 0.80, 12 reviewers would be required with item analyzed as a fixed facet.

Because most manuscript decisions are absolute decisions, the index of dependability is the most important coefficient to interpret in a publication context. However, neither coefficient reached an acceptable level without a very high number of reviewers, which suggests that addressing the issue of reliability may require other changes besides those included here. The number of items and reviewers can be increased, but this is only possible up to a realistic amount (i.e., neither 33 nor 16 reviewers is realistic). One possible alternative is implementing reviewer training to reduce some of the variability. Additionally, other sources of variability may need to be explored and addressed.

**Many-Facet Rasch Measurement**

Research Question 3: Do raw publishability scores versus theta scores predict meaningfully different manuscript decision classifications?

Two multinomial logistic regressions were conducted to predict manuscript decision classifications. The first regression used average raw total score as the predictor variable. The second regression used the publishability measure to predict decision classification.

The first multinomial logistic regression was conducted using average raw total scores. A one unit increase in average raw total was associated with 2.05 (95% CI: 1.72 to 2.43) times the odds of receiving an accept/minor revision decision versus a reject decision (Table 8). A one unit increase in average raw total was associated with 1.40 (95% CI: 1.25 to 1.58) times the odds of receiving a major revision decision versus a reject decision.

Table 8

Multinomial Logistic Regression Using Average Raw Total

| | | Estimate | SE | Wald | df | p-value | OR (95%CI) |
|---|---|---|---|---|---|---|---|
| Accept/Minor Revision | Intercept | -11.31 | 1.39 | 66.15 | 1 | < 0.001 | ----- |
| | Raw Total Score | 0.72 | 0.09 | 66.18 | 1 | < 0.001 | 2.05 (1.72 to 2.43) |
| Major Revision | Intercept | -4.35 | 0.85 | 26.00 | 1 | < 0.001 | ----- |
| | Raw Total Score | 0.34 | 0.06 | 31.36 | 1 | < 0.001 | 1.40 (1.25 to 1.58) |
| Reject | Intercept | ----- | ----- | ----- | ----- | ----- | ----- |
| | Raw Total Score | ----- | ----- | ----- | ----- | ----- | ----- |

*Note*. SE = standard error; df = degrees of freedom; OR = odds ratio; CI = confidence interval

In the multinomial logistic regression, 11.96% of manuscripts were predicted to receive an accept/minor revision decision, 64.78% were predicted to receive a major revision decision, and 23.26% were predicted to be rejected (Table 9). When compared to the actual editor's decisions that were made for the manuscripts, the model correctly classified 30.77% of manuscripts with accept/minor revision decisions, 71.83% of manuscripts with major revision decisions, and 46.81% of manuscripts with reject decisions. The overall percentage of manuscripts that were correctly classified was 55.15%. From these results, the model appears to most accurately classify manuscripts with major revision decisions and is less accurate in classifying manuscripts that receive one of the other two decision categories. See Appendix C for predicted frequencies at each score level. Pseudo R-squared measures were 0.29 (Cox and Snell), 0.34 (Nagelkerke), and 0.17 (McFadden).

Table 9

Classification of Manuscripts Using Average Raw Total

| | Predicted | | | |
|---|---|---|---|---|
| Observed | Accept/Minor Revision | Major Revision | Reject | Percent Correct |
| Accept/Minor Revision | 20 | 45 | 0 | 30.77% |
| Major Revision | 14 | 102 | 26 | 71.83% |
| Reject | 2 | 48 | 44 | 46.81% |
| Overall Percentage | 11.96% | 64.78% | 23.26% | 55.15% |

The second multinomial logistic regression was conducted using the manuscript publishability measure. A one unit increase in the manuscript publishability measure was associated with 2.56 (95% CI: 1.96 to 3.36) times the odds of an accept/minor revision decision versus a reject decision (Table 10). A one unit increase in the manuscript publishability measure was associated with 1.73 (95% CI: 1.40 to 2.13) times the odds of a major revision decision versus a reject decision.

Table 10

Multinomial Logistic Regression Using Publishability Measure

| | | Estimate | SE | Wald | df | p-value | OR (95%CI) |
|---|---|---|---|---|---|---|---|
| Accept/Minor Revision | Intercept | -0.51 | 0.20 | 6.85 | 1 | 0.009 | ----- |
| | Raw Total Score | 0.94 | 0.14 | 46.41 | 1 | < 0.001 | 2.56 (1.96 to 3.36) |
| Major Revision | Intercept | 0.55 | 0.15 | 14.05 | 1 | < 0.001 | ----- |
| | Raw Total Score | 0.55 | 0.11 | 25.76 | 1 | < 0.001 | 1.73 (1.40 to 2.13) |
| Reject | Intercept | ----- | ----- | ----- | ----- | ----- | ----- |
| | Raw Total Score | ----- | ----- | ----- | ----- | ----- | ----- |

*Note*. SE = standard error; df = degrees of freedom; OR = odds ratio; CI = confidence interval

In the multinomial logistic regression using the publishability measure to predict the editor's acceptance decision, 5.98% of manuscripts were predicted to receive an accept/minor revision decision, 73.42% were predicted to receive a major revision decision, and 20.60% were predicted to be rejected (Table 11). When compared to the actual editor's decisions, the model correctly classified 13.85% of manuscripts with accept/minor revision decisions, 78.87% of manuscripts with major revision decisions, and 39.36% of manuscripts with reject decisions. Overall, 52.49% percent of manuscripts were correctly classified when the publishability measure was used. The model most accurately classified manuscripts with major revision decisions and less accurately classified those in the other two decision categories, especially those in the accept/minor revision category. Pseudo R-squared measures were 0.20 (Cox and Snell), 0.22 (Nagelkerke), and 0.10 (McFadden).

Table 11

Classification of Manuscripts Using Publishability Measure

| | Predicted | | | |
|---|---|---|---|---|
| Observed | Accept/Minor Revision | Major Revision | Reject | Percent Correct |
| Accept/Minor Revision | 9 | 54 | 2 | 13.85% |
| Major Revision | 7 | 112 | 23 | 78.87% |
| Reject | 2 | 55 | 37 | 39.36% |
| Overall Percentage | 5.98% | 73.42% | 20.60% | 52.49% |

When comparing the two predictors of decisions, the average raw total score was slightly more accurate at predicting manuscript decision classifications compared to the publishability measure (55.15% correct versus 52.49% correct). While neither method was very accurate at predicting accept/minor revision decisions, the average raw total score was remarkably better than the publishability measure (30.77% versus 13.85%). The percentage of correctly predicted rejection decisions was low for both methods with average raw total score predicting slightly more accurately (46.81% versus 39.36%). Major revision decisions were most accurately predicted by both methods. In this case, the publishability measure was slightly more accurate than the average raw total score (78.87% versus 71.83%). See Appendix D for predicted frequencies at each score level.

These differences in percentage of correctly predicted manuscript decision categories suggest that using manuscript publishability measures may provide meaningfully different manuscript classification decision categorizations compared to using average raw total scores. Of note is that the editor's decisions were made based on raw scores, and publishability measures were not available to the editor for decision

making. The fact that the publishability measure was less accurate in predicting the editor's decisions, which were made based on the raw scores, supports the possibility that using publishability measures would lead to different manuscript classifications. This difference is most pronounced for accept/minor revision decisions where 65 such decisions were observed, 36 such decisions were predicted from average raw total scores, and 18 such decisions were predicted from publishability measures. In these cases, the publishability measures likely were lower than the threshold for this category, and the manuscripts would have been classified into another decision category. Similarly, for the rejection category, some manuscripts likely had higher publishability measures than the threshold for the rejection category and were classified into a different category. The results seen here may not reach the strength of such differences that would have been seen if decisions had actually been based on publishability measures. Pseudo R-squared values for the average raw total score model were higher than those for the publishability measure model, which may support the idea that the two scoring methods predict different manuscript decision classifications.

Figure 17 displays the association between average raw total scores and publishability measures. As seen in this figure, the publishability measure increases as the average raw total score increases. The Spearman rank-order correlation of the two is 0.8615, and the Pearson correlation is 0.8564, suggesting a strong association between the two measures of manuscript quality and a shared 73% of variance. In general, rejection decisions are associated with lower average raw total scores and lower publishability measures. Accept/minor revision decisions are associated with higher

average raw scores and higher publishability measures. Major revision decisions occur most often in the middle of the average raw total and publishability measure ranges but are seen at all ranges. This indicates that decision categories are not clear cut, even with the publishability measure after adjustment for reviewer severity.

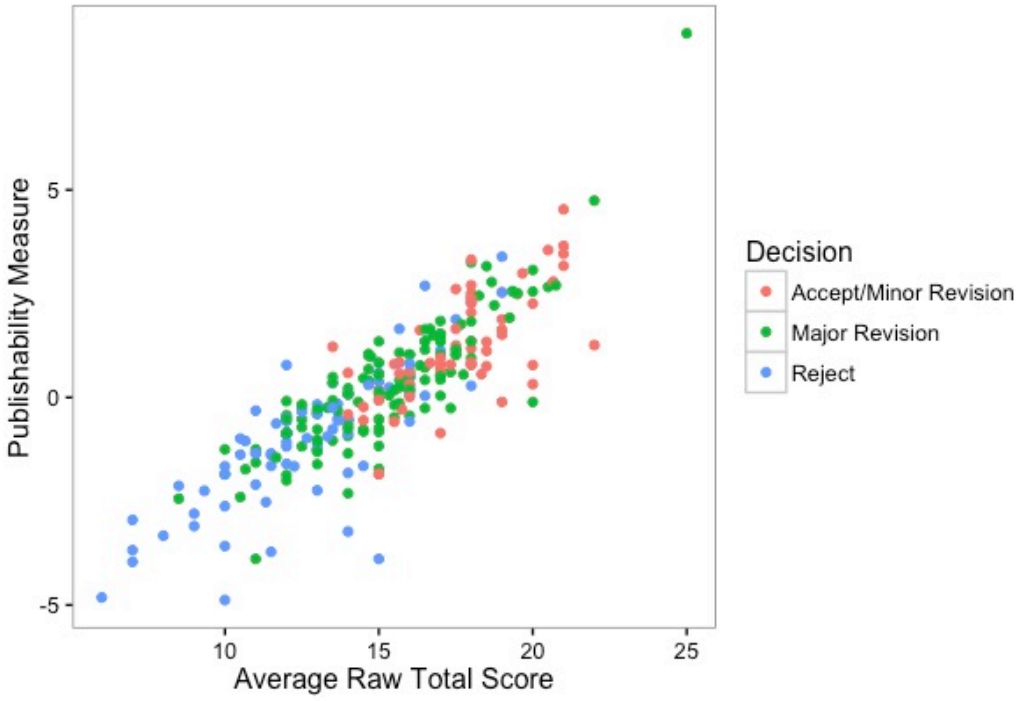


Figure 17. Comparison of Average Raw Scores and Publishability Measures.

The polyserial correlation between manuscript decision and average raw total score was computed as a validity coefficient. The correlation value of -0.60 suggests a moderate negative association between raw scores and manuscript decision categories, meaning that as raw scores increase, a decision of rejection is less likely. This is consistent with the findings of the regression analysis.

Research Question 4: How closely do ranks of the severity measure from each reviewer in a Many-Facet Rasch Measurement analysis compare to ranks of reviewers using average raw ratings from each reviewer?

The Spearman rank-order correlation for the reviewers' average raw ratings with their severity measures from the Many-Facet Rasch Measurement analysis was -0.6083. This can be interpreted as a moderate negative correlation that indicates high average raw ratings are associated with low rater severity rankings (Figure 18). The Pearson correlation was -0.6306 and similarly suggests that high average raw ratings are associated with low reviewer severity. However, both of these correlations indicate that average raw ratings and the reviewer severity measure share only approximately 37% of their variance.

From this analysis, average raw ratings and the reviewer severity measure are moderately correlated but also exhibit differences. After controlling for differences in manuscript publishability quality, the ranks of reviewers are somewhat different. Without the effects of manuscript quality, a clearer picture of reviewer severity is available.

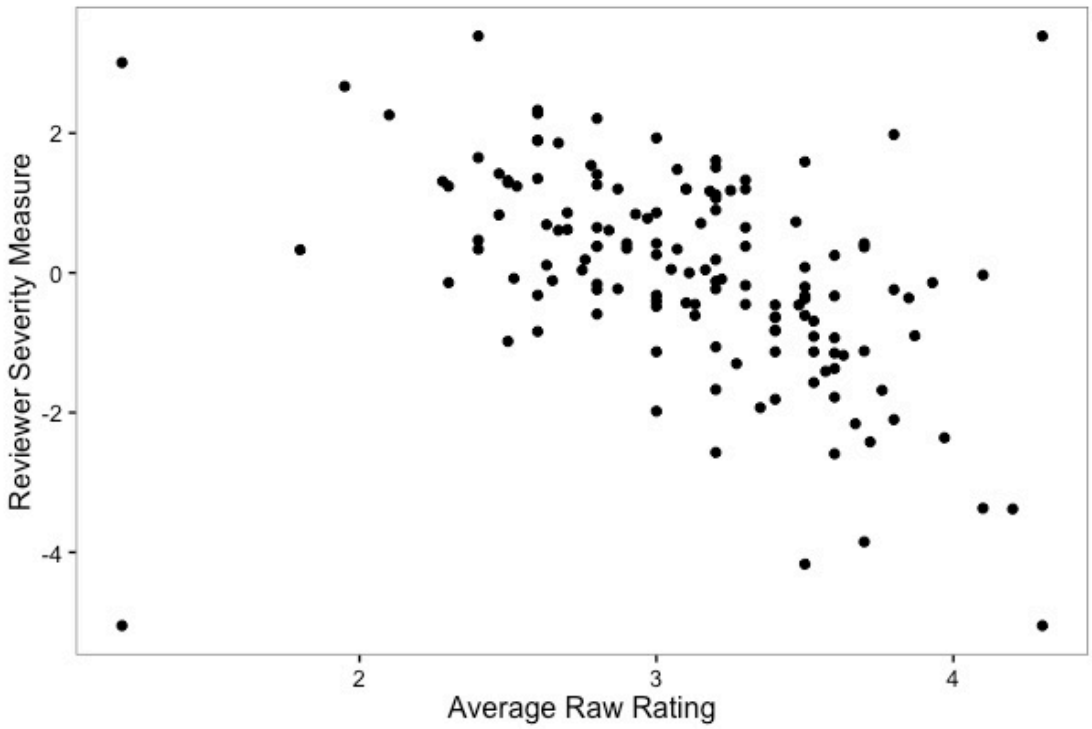


Figure 18. Comparison of Average Raw Ratings and Reviewer Severity Measures.

# CHAPTER V

# DISCUSSION

The purpose of this study was to use Generalizability Theory and Many-Facet Rasch Measurement to examine the effects of reviewer severity on the ratings and decisions made during the peer review of scientific manuscripts. In this section, the findings from the study are discussed and their implications are explored. Additionally, future directions for this line of research are proposed.

## Overview of the Study

### Data

Deidentified peer reviews ($N = 635$) of 301 scientific manuscripts were included in the analyses. Peer reviewers used five items, each with a five-point rating scale, to rate each manuscript according to its publishability. The five publishability criteria were Novelty, Clinical Impact, Scientific Impact, Definitive, and Interesting to Specialty. Publication decisions for each manuscript also were included in the analyses.

### Generalizability Theory Analysis

A Generalizability Theory analysis was conducted with a two-facet, partially nested design. In this design, manuscript served as the object of measurement. Included facets were items and reviewers nested within manuscript. This design allowed calculation of variance components for manuscripts, items, reviewers nested within manuscript, the manuscript-by-item interaction, and the three-way interaction between

manuscript, item, and reviewer plus remaining sources of unmeasured systematic and unsystematic variation.

**Decision Studies**

The variance components produced in the Generalizability Study were used in Decisions Studies to estimate variance components for several combinations of hypothetical numbers of reviewers and items. For each combination, a generalizability coefficient and index of dependability were estimated. The increases in these coefficients were considered in determining the appropriate study design for obtaining results reliable for decision making. An additional analysis was conducted with items set as a fixed facet.

**Many-Facet Rasch Measurement Analysis**

Many-Facet Rasch Measurement analysis was undertaken to produce reviewer severity measures and manuscript publishability measures corrected for reviewer severity. Model infit and outfit were assessed, as well as discrimination parameters. The performance of each item of the five-item scale also was evaluated. Finally, the data were reanalyzed without the reviewer facet to examine the effect of this facet on the ability to detect differences in manuscripts.

**Application to Research Questions**

The above methods and additional analyses were used to address the four research questions. The Generalizability Theory analysis results were used for Research Question 1 and Research Question 2. The Many-Facet Rasch Measurement results were used for Research Question 3 and Research Question 4. Additional multinomial logistic

regression and correlation analyses also were used to address Research Question 3 and Research Question 4.

## Generalizability Theory Findings and Interpretations

The results of the Generalizability Theory analysis revealed that reviewers nested within manuscript account for 35.48% of the variance in publishability scores. This facet had the largest variance component out of all sources of variability. Manuscript, the object of measurement, accounted for only 12.21% of the total variance in publishability scores. Items accounted for 15.22% of the total variance. An additional 3.54% of variance was accounted for by the manuscript-by-item interaction, and 33.55% was accounted for by the three-way interaction of manuscripts, reviewers, and items plus other sources of error.

With the current number of reviewers and items, the generalizability coefficient was 0.3590, and the index of dependability was 0.3295. Because these values are too low for the reviews' scores to be considered reliable for decision making, both the number of reviewers and the number of items were increased in alternative Decision Studies. In the Decision Studies, increasing the number of reviewers and the number of items improved both the generalizability coefficient and the index of dependability. However, to reach an acceptable level of 0.80 for the generalizability coefficient, seven items and 16 reviewers would be required. To increase the index of dependability to 0.80, 10 items and 33 reviewers would be required. Although these increases are hypothetically possible, they are not realistic in real world circumstances.

To explore other analytical means of improving score reliability, the item facet was fixed, which assumed that the differences between items are not due to error. This analysis also required the assumption that the five items were not random samples of publishability questions but fully encompass the construct of publishability. In this case, the generalizability coefficient and index of dependability both improved to 0.4077. For the coefficients to reach 0.80, 12 reviewers would be needed, which would not be reasonable.

When applied to Research Question 1, the above results directly address the proportion of variance attributable to reviewer nested within manuscript variation (35.48%). This was the largest variance component of those analyzed. Manuscripts (12.21%), items (15.22%), and the manuscript-by-item interaction (3.54%) accounted for much smaller proportions of the variance in observed scores. The three-way interaction confounded with the item-by-reviewer interaction and other sources or error accounted for the second largest proportion of total variance, 33.55%. Because the reviewer nested within manuscript variance component is larger than the others, the reviewer variation can be thought to greatly contribute to the variation in observed scores. This variation is considered error and accounts for the low dependability observed in the study.

When applied to Research Question 2, the results indicate that the conditions of measurement should be changed, but the extent to which they should be changed is too great to be realistically possible. The original study design's generalizability coefficient ($\rho^2 = 0.3590$) and index of dependability ($\Phi = 0.3259$) were too low for use in making publication decisions. In alternative Decision Studies, the generalizability coefficient

reached the acceptable level of 0.80 with seven items and 16 reviewers, and the index of dependability reached the acceptable level of 0.80 with 10 items and 33 reviewers. Although these increased numbers of items and reviewers could raise the two coefficients to levels suitable for decision making, such high numbers of items and reviewers are not realistic. Additional changes besides increasing the number of reviewers and items may be necessary, and other sources of variability may need to be considered. Such changes could include reviewer training or modifications to the scoring rubric.

The sizeable reviewer nested within manuscript effect, which accounted for 35.48% of the variance in observed scores, implies that reviewer severity is a major barrier to obtaining comparable manuscript publishability scores. The large reviewer nested within manuscript variance component was different than expected because, ideally, the object of measurement, manuscript, variance component should be the largest (Shavelson & Webb, 1991). Variability in the manuscripts under review should contribute more than the reviewers to the observed score variability. Because this variance component includes the reviewer-by-manuscript interaction confounded with the reviewer effect, some of this variability could be due to differences in reviewer behavior across manuscripts. The reviewer-by-item interaction was not measurable in this design, but those effects are included in the large residual variance component. This variance component, which also includes the three-way interaction of manuscripts, reviewers, and items accounts for 33.55% of the total variance. This is cause for concern because after accounting for variability from the other facets and interactions, a large portion of the variance is not explained and must be attributable to other sources, and a great amount of

undifferentiated error remains to be understood and improved. The manuscript-by-item interaction variance component was small and not a cause for concern.

When the item facet was fixed, the variability associated with this facet was removed from the analysis. This increased both the generalizability coefficient and the index of dependability. The number of reviewers required for these coefficients to reach 0.80 also was reduced. Based on these results, removing the variability associated with items seems to improve the reliability of observed manuscript scores, but the number of required reviewers is not reasonable in a real world setting. This finding is not unexpected given the moderate item facet effect (15.22%) and the manuscript-by-item interaction effect (3.54%). An additional concern it that fixing the item facet assumes that the five included items are the only publishability items that could be used in the universe of generalization (Brennan, 2001; Shavelson & Webb, 1991). Because publishability has many aspects, this assumption does not seem reasonable when realistically considering the construct. Without further research, no evidence is available to support the use of these items to represent all aspects of publishability.

**Many-Facet Rasch Measurement Findings and Interpretations**

The results from the Many-Facet Rasch Measurement analysis provided information on each of the facets and the fit of the model. The observed average manuscript score was 3.00 (SD = 0.60), and the fair average, produced after adjustment for reviewer severity and item difficulty, was slightly higher at 3.07 (SD = 0.65). This finding suggests that manuscript scores may improve by a very small amount, if scores were based on the average reviewer and items of average difficulty. The publishability

(theta) measure had an average of 0.14 (SD = 1.66), and the included manuscripts ranged in publishability levels from -4.88 to 8.77.

The fit of the model to the included manuscripts was not consistent with 45.18% of manuscripts exhibiting infit problems and 45.85% exhibiting outfit problems. The majority of these fit problems were due to too little variation in scores (30.23% with infit < 0.5). The high percentage of manuscripts with infit statistics below the acceptable level suggests that manuscripts were rated too consistently (Linacre, 2002; Myford & Wolfe, 2004). Perhaps manuscripts were consistently rated high or low with little variation between the two extremes. Infit statistics that fell above the threshold suggest too much variation and scores that were inconsistent and erratic. Outfit statistics that fell outside the acceptable range reveal instances of too little variation or too much variation from typically consistent manuscripts. Problems with discrimination occurred in 47.50% of manuscripts, suggesting that manuscript scores do not do well at distinguishing reviewers from each other or items from each other. Despite these issues, the manuscripts were reliably different from each other (reliability of separation index = 0.89). However, the number of discrimination and fit problems suggests that manuscript scores should not be used to distinguish reviewers from each other or items from each other (Linacre, 2002).

One possible explanation for the misfit of the manuscript scores to the model is the absence of interaction terms in the model. The manuscript-by-item interaction or manuscript-by-reviewer interaction cannot be assessed in this case, which means any variability due to these interactions will not be represented in the model. If such

interaction effects exist, the model will fit less well than it could have had these effects been included.

The mean reviewer severity measure was 0.00 (SD = 1.43) with a range from -5.05 to 3.39. This range indicates reviewers varied in their severity when reviewing manuscripts. The observed average rating reviewers provided was 3.11 (SD = 0.51) and the fair average was 3.06 (SD = 0.58) after accounting for manuscript quality and item difficulty. This difference is very small and does not suggest much change in ratings after adjustment.

Like the manuscript facet, the reviewer facet showed problems with fit to the model. Of the included reviewers, 41.30% exhibited infit problems, and 40.58% exhibited outfit problems. The high number of reviewers with infit statistics above (26.09%) or below (15.22%) the acceptable level suggests that some reviewers were too consistent in their ratings, and some reviewers were too inconsistent in their ratings (Linacre, 2002; Myford & Wolfe, 2004). Outfit statistics commonly fell outside the acceptable range, indicating instances of too little variation (15.22%) or too much variation (25.36%) from typically consistent reviewers. Discrimination problems were present in 42.75% of reviewers, suggesting that reviewers do not discriminate well among manuscripts of different quality levels. An additional concern was identified by the reliability of separation index (0.90), which revealed that reviewers are reliably different from each other and do not provide comparable ratings on the same manuscripts (Linacre, 2002). However, this also means that the pool of data is large enough to allow reviewer severity measures to be well estimated and, therefore, corrected for in the

model. The large standard deviation of the severity measure is another sign that reviewers differ in severity.

The lack of interaction terms may also explain some of the model misfit seen with the reviewer facet. A reviewer-by-item interaction could provide insight into whether reviewers' severity levels differ across items. However, the model will fit less well, if this interaction is occurring but not included. The lack of similarity across reviewers may be due to differences in the application of the rating scale or differences in understanding of manuscript quality. The discrimination problems seem to support this notion.

The five items ranged in difficulty from -1.35 to 0.87 with the Interesting to Specialty item ranked as least difficult and the Definitive item ranked as most difficult. No infit or outfit problems were identified for any of the items, suggesting a good fit with the model and supporting the use of the rating scale model. Discrimination values were acceptable for all items, and the items were reliably different from each other (reliability of separation index = 0.99). Additionally, each of the items performed well with all response categories used and response categories ordered correctly in difficulty. The absence of infit and outfit problems and discrimination problems shows that the model fits well, and scores on the items reflect scores on the overall scale (Linacre, 2002). The flat item information curve reflects the polytomous nature of the items and the fact that the items were not written to target levels associated with a cut score.

An additional analysis was conducted without the reviewer facet. In this analysis, the manuscript reliability of separation index decreased to 0.83 when it had been 0.89

with the reviewer facet included. This suggests that the reviewer facet does contribute to the detection of differences in manuscripts.

The results of the item analysis support the use of the rating scale model and of Many-Facet Rasch Measurement in general (Andrich, 1978, Linacre, 2000, Wright, 1998). While the other facets showed signs of trouble in model fit, the item facet fit very well, and the items themselves performed as expected. Other factors may play a role in the misfit of manuscripts and reviewers, but this method still provides useful information to the analysis. This is consistent with the results from the Generalizability Theory analysis where the item facet was a small source of variance.

**Comparison of Raw Ratings and Many-Facet Rasch Measurement Results**

To determine whether raw publishability scores versus theta scores predicted meaningfully different manuscript decision classifications, Research Question 3, average raw total scores and manuscript publishability measures (theta) were used. When the average raw total score was used to predict manuscript decision category, the overall percentage of manuscripts that were correctly classified using the average raw total score was 55.15%, and the model appeared most accurate at classifying manuscripts with major revision decisions and less accurate at classifying manuscripts in the other decision categories. Using the manuscript publishability measure (theta), the percentage of manuscripts that were correctly classified when the publishability measure was used was 52.49%, and the model appeared most accurate at classifying manuscripts with major revision decisions and less accurate at classifying those in the other categories, particularly the accept/minor revision category.

The average raw total score and the publishability measure have a Spearman rank-order correlation of 0.8615 and a Pearson correlation of 0.8564, indicating that the two measures are strongly related and share 73% of their variance. Taken with the results from regression analysis, this finding suggests that the two measures are very similar but have some differences that could lead to meaningful differences in manuscript classification. The manuscript publishability measure did not correctly classify as many manuscripts as the average raw total score. Because the decisions included in this dataset were based on the raw scores, the publishability measure should correctly classify fewer manuscripts than the average raw total scores, if there truly are differences in classifications.

Reviewers' severity measures and their average raw ratings were used in Research Question 4 to examine how closely ranks of the severity measure from each reviewer in Many-Facet Rasch Measurement compared to ranks of reviewers using their average raw ratings. The reviewers' average raw ratings and the reviewers' severity measures have a Spearman rank-order correlation of -0.6083 and a Pearson correlation of -0.6306. These correlations indicate that high average raw ratings are associated with low reviewer severity. These are moderate correlations that correspond to approximately 37% shared variance. These differences likely are attributable to the adjustment for manuscript quality in the severity measure. When the effects of manuscript quality are accounted for, the severity measure is able to provide an improved assessment of reviewers' tendencies.

## Summary of Similarities and Differences Between Methods

Generalizability Theory and Many-Facet Rasch Measurement both provide information on manuscripts, reviewers, and items in the analysis of reviews of scientific manuscripts. The main result of Generalizability Theory analysis is the amount of variance that can be attributed to the object of measurement and each facet in the study. In the current analysis, most of the variability in observed scores could be attributed to reviewers nested within manuscripts. The manuscripts did not contribute as much to the variability in observed scores, but unaccounted for sources of variability contributed greatly to the variance in scores. Many-Facet Rasch Measurement does not provide information on the proportion of variability contributed by each facet, but it does provide detailed information for each facet, including measures that are adjusted for the other facets.

Decision Studies provide information about how reliable the observed scores are for making decisions and whether changes should be made to the study design (Cronbach, et al.,1972; Shavelson & Webb, 1991). The results of the current study suggest that the scores may not be reliable for decision making, and changes to the rating process should be made. However, the extent of the changes may be beyond the scope of the study design. Many-Facet Rasch Measurement calculates a different type of reliability that describes the extent to which the elements of each facet are reliably different. In this study, the reviewers were found to be reliably different in their reviews of the same manuscripts. These results correspond to the large reviewer variance component from the Generalizability Theory analysis. The manuscript and item facets

also were reliably different in the Many-Facet Rasch Measurement analysis, similar to the variance components from Generalizability Theory. Although the reliability calculations from the two methods are different, they both suggest changes are needed, especially to the reviewer facet, which appears to have too much variability.

An advantage of Generalizability Theory over Many-Facet Rasch measurement is that it allows for the inclusion of interactions between facets, making it possible to assess whether one facet differs across levels of another facet. This is not possible with Many-Facet Rasch Measurement and not only prevents understanding of potential interactions but possibly leads to model misfit. In the current study, the large interactions revealed in the Generalizability Theory analysis are indicative of problems that lead to misfit and low discrimination. The Generalizability Theory analysis provided insight into the problems that Many-Facet Rasch Measurement indicated but did not have means to explain.

In Generalizability Theory, the presence of error can be assessed with the residual variance component (Brennan, 2001). When this variance component is large, sources of unmeasured variability likely have influenced the observed scores. This was the case for the current study in which the residual variance accounted for 33.55% of the total variance. In Many-Facet Rasch Measurement, the reliability of separation index provides a sense of the amount of error. The high indices for each facet in the analysis suggest that the observed differences are due to reliable differences in manuscripts, reviewers, and items.

Perhaps the greatest advantage of Many-Facet Rasch Measurement is its adjustment for other facets in the estimation of theta. When computing reviewer severity,

the analysis is adjusted for manuscript quality and item difficulty. When computing manuscript publishability, the analysis is adjusted for reviewer severity and item difficulty. In the analysis of reviewer severity, this assists with understanding reviewers' tendencies regardless of the specific manuscripts they reviewed. Generalizability Theory does not have this type of adjustment and does not provide sophisticated measures of reviewer severity or manuscript publishability.

When manuscript decision classifications predicted by average raw total scores and publishability measures were compared, differences in classifications were seen. The adjustment for reviewer severity and item difficulty made some difference in the publishability measure that led to different publication decision categories. While the average raw scores and publishability measures were highly and positively correlated and explained some but not all variability, all manuscripts did not fall into the same categories. Based on these results, the publishability measure may provide a clearer picture of manuscript quality.

Similarly, a comparison of the reviewer severity measure and average raw ratings from reviewers demonstrated differences in rankings of reviewers. Higher average raw ratings were associated with lower rater severity, but this correlation was only moderate. In the case of reviewer severity, controlling for manuscript quality seems to be very important in getting an accurate picture of severity. This seems to be a valuable advantage that Many-Facet Rasch Measurement has over Generalizability Theory. Additionally, Generalizability Theory describes differences in facets as error variance

while Many-Facet Rasch Measurement attempts to correct for those differences, thus removing their contribution to error.

The results of the current study concur with previous work (Kim & Wilson, 2009; MacMillan, 2000; Sudweeks et al., 2005). These studies found that aspects of the two methods provided comparable, though not identical, conclusions. While the two methods of assessing rater effects have many differences, they each have their own advantages. The variance components, interactions, and reliability assessment from Generalizability Theory and the detailed facet information and adjusted theta scores produced from Many-Facet Rasch Measurement all are important elements for assessing rater effects. Because neither method is clearly superior to the other, using the two methods as complements to each other would produce the most comprehensive understanding of how a measure is functioning.

## Implications for the Peer Review Process

The results of this study have many implications for the peer review process. One of the most important findings from the study is that reviewers were not consistent in their assignment of scores, and they did not reliability score manuscripts in a similar manner. This variability likely affects scientific knowledge when manuscripts are either published or not published based on these reviews.

The analysis of this dataset revealed that publication decisions cannot be reliably made from these manuscript scores. While the generalizability of the results is unknown, many journals likely could benefit from improved methods of peer review. These results show that reviewers exhibit a very large amount of variability in their reviews. This

variability is so great that the Decision Studies indicated it could not be overcome with a realistic number of additional reviewers and items. Additional solutions such as reviewer training and rubric improvement could help reduce some of the reviewer variability. Peer reviewers could be trained on aspects of manuscript quality and how to use the five items to assess quality. Journal editorial staff may benefit from knowledge that such problems exist, and, in turn, can be prepared to address or prevent problems with reviewer severity.

By strictly interpreting the statistical findings of the analyses, the reliability of the reviewers' scores of manuscripts appears low. However, an alternative interpretation suggests that the low reliability may be intentional and even desirable. In some cases, journals select reviewers for a manuscript in a way that maximizes variability. For example, the panel of reviewers for a particular manuscript may be composed of both content experts and methodology experts. This purposeful selection of a diverse panel of reviewers ensures that all important aspects of a strong manuscript are scrutinized, but the level of examination of each point may differ by reviewer. The aspects of a manuscript that are most important to and most scrutinized by each of these types of reviewers likely vary. For example, a content expert may be most concerned with the clinical impact of a manuscript, and a methodology expert may be most concerned with the scientific impact or the definitiveness of the research. If this is the case, some manuscript rating items are more important to some reviewers than to others.

Under these circumstances, some reviewer variance would be expected in a Generalizability Theory analysis. Reviewers would not give the same consideration to the same items on the rating scale and would, therefore, seem to review the manuscripts

in a different manner. If such intentional variability is considered in the analysis, a large reviewer variance component is not as problematic as would be the case if all reviewers where expected to behave similarly. In fact, this possibility could be considered in the analysis with certain items belonging to certain components of publishability. Such a structure would allow appropriate variance to be accounted for and understood while separating it from unwanted variance.

## Application of Methods to Dataset

Generalizability Theory and Many-Facet Rasch Measurement analyses traditionally have been conducted on balanced datasets that have not featured nesting (Cronbach et al., 1972; MacMillan, 2000; Shavelson & Webb, 1991). Many-Facet Rasch Measurement methods are best applied to data with strong connectivity among raters (Linacre, 2014). The data used in the current analyses were unbalanced and included nesting. The data were not as strongly connected as data that have been used in many analyses. Despite these concerns, the analyses were possible and produced usable results. Applying these methods to imperfect data in fields beyond those for which they were developed improves understanding of reviewer effects despite the limitations of the data and serves as a starting point for understanding phenomena previously unstudied. The successful application of these methods to this real world dataset suggests potential for use of Generalizability Theory and Many-Facet Rasch Measurement with other similar datasets.

## Limitations

While the proposed study has the strengths of applying established methods to a new field and carrying out a comparison of the results, the design was not without limitations. In the Generalizability Theory study, the fact that the same individuals reviewed more than one manuscript could not be considered under the current design. This design assumes that each reviewer only reviewed one manuscript. With this limitation, examining a reviewer's behavior over multiple manuscripts was not possible.

Another problem with the nested design is that the reviewer effect could not be examined separately. The reviewer effect was confounded with the reviewer-by-manuscript interaction, precluding the possibility of determining whether reviewers differed in severity alone or whether some reviewers were harsher for some manuscripts. For this reason, conclusions about reviewer severity are somewhat ambiguous. The design also makes examining the item-by-reviewer interaction impossible, limiting the information that can be learned from the analysis.

Furthermore, the data used in the study is limited to that which was available in the review database. There was no control over which reviewers rated which manuscripts or how many reviews occurred and were included in the dataset. The results may not be generalizable to other journals where a different process and different reviewers are used. However, these analyses do provide insight into how such studies perform with real-world, imperfect datasets.

When examining decisions based on raw publishability scores versus theta scores (Research Question 3), the manuscript decisions were based on the existing data and not

on the information generated from the Many-Facet Rasch Measurement analysis. Editors were not actually able to use theta scores to make new decisions. Consequently, potential differences due to scoring likely were minimized and potentially underestimated, and conclusions about the influence of results on decision making were constrained.

## Future Directions

Future research in this area would benefit from data collection efforts designed specifically for the purpose of assessing rater effects. A fully crossed study without nesting would allow assessment of the reviewer facet without confounding it with the manuscript facet. A reviewer-by-manuscript interaction and a reviewer-by-item interaction also could be calculated in this study design. Other facets such as study type (e.g., randomized controlled trial, cross-sectional, systematic review, etc.) could be included to explore potential effects beyond those examined in the current study. Further analyses could include hierarchical nesting by manuscript type (e.g., technical manuscripts versus applied manuscripts).

Additionally, a reviewer training experiment could be conducted by introducing training methods to a group of reviewers and comparing their performance to reviewers who have not received the training. Ideas for training could be gathered by interviewing misfitting reviewers about their reviewing style. Variability in reviewer type (e.g., content versus methodology experts) could be considered and included and addressed in the manuscript scoring system and analyses. Finally, new items could be tested to determine whether they improve reliability of publishability scores.

## Conclusions

This study applied Generalizability Theory and Many-Facet Rasch Measurement to the field of peer review. The study's findings indicate that reviewers are inconsistent in their reviews of manuscripts, both as individuals and as a group. The advantages of both Generalizability Theory and Many-Facet Rasch Measurement contributed to the results of the study, and both methods were useful in understanding the reviewer data. The use of these methods in peer review of scientific manuscripts will increase the capacity for more fair and accurate rating methods in this field. Although the study has its limitations, the results have the potential to bring positive change to peer review.

# REFERENCES


Andrich, D. (1978). A rating scale formulation for ordered response categories. *Psychometrika,* 43, 561-573.

Benos, D. J., Bashari, E., Chaves, J. M., Gaggar, A., Kapoor, N., LaFrance, M., ... Zotov, A. (2007). The ups and downs of peer review. *Advances in Physiology Education, 31*(2), 145-152.

Blackburn, J. L., & Hakel, M. D. (2006). An examination of sources of peer-review bias. *Psychological Science, 17*(5), 378-382.

Bornmann, L., Mutz, R., & Daniel, H. D. (2010). A reliability-generalization study of journal peer reviews: a multilevel meta-analysis of inter-rater reliability and its determinants. *PloS one*, *5*(12), e14331.

Brennan, R. L. (1983). *Elements of generalizability theory*. American College Testing Program.

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement*, *24*(4), 339-353.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Brennan, R. L., & Johnson, E. G. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice*, *14*(4), 9-12.

Brennan, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement*, *14*(3), 277-289.

Callaham, M. L., Wears, R. L., Weber, E. J., Barton, C., & Young, G. (1998). Positive-outcome bias and other limitations in the outcome of research abstracts submitted to a scientific meeting. *JAMA, 280*(3), 254-257.

Cho, M. K., Justice, A. C., Winker, M. A., Berlin, J. A., Waeckerle, J. F., Callaham, M. L., ... the PEER Investigators. (1998). Masking author identity in peer review: What factors influence masking success?. *JAMA, 280*(3), 243-245.

Cicchetti, D. V., & Conn, H. O. (1976). A statistical analysis of reviewer agreement and bias in evaluating medical abstracts. *The Yale Journal of Biology and Medicine*, *49*(4), 373-383.

Colliver, J. A., Verhulst, S. J., Williams, R. G., & Norcini, J. J. (1989). Reliability of performance on standardized patient cases: A comparison of consistency measures based on generalizability theory. *Teaching and Learning in Medicine: An International Journal*, *1*(1), 31-37.

Congdon, P. J., & McQueen, J. (2000). The Stability of Rater Severity in Large‐Scale Assessment Programs. *Journal of Educational Measurement*, *37*(2), 163-178.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley & Sons.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281.

de Gruijter, D. N. (1984). Two simple models for rater effects. *Applied Psychological Measurement, 8*(2), 213-218.

Du, Y., & Brown, W. L. (2000). Raters and single prompt-to-prompt equating using the facets model in a writing performance assessment. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice* (Vol. 5). Stamford, CT: Ablex Publishing Corporation.

Du, Y., Wright, B. D., & Brown, W. L. (1996, April). *Differential facet functioning detection in direct writing assessment*. Paper presented at the Annual Conference of American Educational Research Association, New York, NY.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, *4*(4), 289-303.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155-185.

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Section H). Strasbourg, France: Council of Europe/Language Policy Division.

Emerson, G. B., Warme, W. J., Wolf, F. M., Heckman, J. D., Brand, R. A., & Leopold, S. S. (2010). Testing for the presence of positive-outcome bias in peer review: A randomized controlled trial. *Archives of Internal Medicine, 170*(21), 1934-1939.

Engelhard, G. (1992). The measurement of writing ability with a Many-Faceted Rasch Model. *Applied Measurement in Education*, *5*(3), 171-191.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a Many-Faceted Rasch Model. *Journal of Educational Measurement*, *31*(2), 93-112.

Engelhard, G. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement, 33*(1), 56-70.

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York: Routledge Academic.

Farrokhi, F., Esfandiari, R., & Schaefer, E. (2012). A many-facet Rasch measurement of differential rater severity/leniency in three types of assessment. *Japan Association for Language Teaching Journal*, *34*(1), 79-101.

García, J. A., Rodriguez-Sánchez, R., & Fdez-Valdivia, J. (2015). The principal-agent problem in peer review. *Journal of the Association for Information Science and Technology, 66*(2), 297-308.

Gingerich, A., Regehr, G., & Eva, K. W. (2011). Rater-based assessments as social judgments: rethinking the etiology of rater errors. *Academic Medicine*, *86*(10), S1-S7.

Godlee, F., Gale, C. R., & Martyn, C. N. (1998). Effect on the quality of peer review of blinding reviewers and asking them to sign their reports: A randomized controlled trial. *JAMA, 280*(3), 237-240.

Grainger, D. W. (2007). Peer review as professional responsibility: A quality control system only as good as the participants. *Biomaterials, 28*(34), 5199-5203.

Holzbach, R. L. (January 01, 1978). Rater bias in performance ratings: Superior, self-, and peer ratings. *Journal of Applied Psychology, 63,* 5, 579-588.

Houston, W. M., Raymond, M. R., & Svec, J. C. (1991). Adjustments for rater effects in performance assessment. *Applied Psychological Measurement, 15*(4), 409-421.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it?. *Psychological Methods*, *5*(1), 64.

Jefferson, T., Rudin, M., Brodney Folse, S., & Davidoff, F. (2007). Editorial peer review for improving the quality of reports of biomedical studies. *Cochrane Database of Systematic Reviews*. doi: 10.1002/14651858.MR000016.pub3

Jefferson, T., Wager, E., & Davidoff, F. (2002). Measuring the quality of editorial peer review. *JAMA, 287*(21), 2786-2790.

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, *26*(4), 485-505.

Kane, J. S., Bernardin, H. J., Villanova, P., & Peyrefitte, J. (1995). Stability of rater leniency: Three studies. *Academy of Management Journal*, *38*(4), 1036-1051.

Kane, M. (1999, April). *The role of generalizability in validity*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Kane, M. T. (1982). A sampling model for validity. Applied *Psychological Measurement, 6*(2), 125-160.

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), Educational measurement (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1-73.

Kassirer, J. P., & Campion, E. W. (1994). Peer review: Crude and understudied, but indispensable. *JAMA, 272*(2), 96-97.

Kim, S. C., & Wilson, M. (2009). A comparative analysis of the ratings in performance assessment using Generalizability Theory and the Many-Faceted Rasch Model. *Journal of Applied Measurement*, *10*(4), 403-423.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*(1), 3-31.

Lakes, K. D., & Hoyt, W. T. (2008). What sources contribute to variance in observer ratings? Using generalizability theory to assess construct validity of psychological measures. *Infant and Child Development, 17*(3), 269-284.

Langfeldt, L. (2006). The policy challenges of peer review: managing bias, conflict of interests and interdisciplinary assessments. *Research Evaluation, 15*(1), 31-41.

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology, 64*(1), 2-17.

Lin, C. K. (2014). *Issues and challenges in current generalizability theory applications in rated measurement* (Doctoral dissertation). University of Illinois at Urbana-Champaign, Champaign, IL.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. (1993). Generalizability Theory and Many-facet Rasch Measurement, presented at the 1993 Annual Meeting of the American Educational Research Association, Atlanta, GA, 1993. Washington, DC: American Educational Research Association.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), 328.

Linacre, J.M. (2000). Comparing "Partial Credit Models" (PCM) and "Rating Scale Models" (RSM). *Rasch Measurement Transactions, 14*(3), 768.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878.

Linacre, J. M. (2014). A User's Guide to FACETS Rasch-Model Computer Programs Program Manual, version 3.71.4. Beaverton, Oregon: Winsteps.com

Linacre, J. M., Engelhard, G., Tatum, D. S., & Myford, C. M. (1994). Measurement with judges: Many-faceted conjoint measurement. *International Journal of Educational Research*, *21*(6), 569-577.

Link, A. M. (1998). US and non-US submissions: An analysis of reviewer bias. *JAMA, 280*(3), 246-247.

Lock, S. (1985). *A difficult balance: Editorial peer review in medicine*. London: Nuffield Provincials Hospital Trust.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley Pub. Co.

Luecht, R. M. (1989). *A comparison of applied methods for estimating variance components under large, random-effects designs with unbalanced data* (Order No. 9002907). Available from ProQuest Dissertations & Theses Global. (303742547).

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54-71.

Lunz, M. E., & Stahl, J. A. (1993). The effect of rater severity on person ability measure: A Rasch model analysis. *American Journal of Occupational Therapy*, *47*(4), 311-317.

MacMillan, P. D. (2000). Classical, generalizability, and multifaceted Rasch detection of interrater variability in large, sparse data sets. *The Journal of Experimental Education*, *68*(2), 167-190.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1*(2), 161-175.

Marsh, H. W., & Ball, S. (1981). Interjudgmental reliability of reviews for the Journal of Educational Psychology. *Journal of Educational Psychology, 73*(6), 872-880.

Marsh, H. W., & Ball, S. (2014). The peer review process used to evaluate manuscripts submitted to academic journals. *The Journal of Experimental Education, 57*(2), 151-169.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrik*a, 47, 149-174.

McNutt, R. A., Evans, A. T., Fletcher, R. H., & Fletcher, S. W. (1990). The effects of blinding on the quality of peer review: a randomized trial. *JAMA, 263*(10), 1371-1376.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). Washington, DC: American Council on Education / Macmillan.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386-422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189-227.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, *15*(2), 158-180.

Opthof, T., Coronel, R., & Janse, M. J. (2002). The significance of the peer review process against the background of bias: Priority ratings of reviewers and editors and the prediction of citation, the role of geographical bias. *Cardiovascular Research, 56*(3), 339-346.

Prieto, G., & Nieto, E. (2014). Analysis of rater severity on written expression exam using Many Faceted Rasch Measurement. *Psicológica*, *35*(2), 385-397.

Raymond, M. R., Harik, P., & Clauser, B. E. (2011). The impact of statistically adjusting for rater effects on conditional standard errors of performance ratings. Applied Psychological Measurement, 35(3), 235-246.

Raymond, M. R., & Houston, W. M. (1990). Detecting and Correcting for Rater Effects in Performance Assessment. (ACT Research Report Series 90-14). Iowa City, IA: The American College Testing Program.

Raymond, M. R., & Viswesvaran, C. (1993). Least squares models to correct for rater effects in performance assessment. *Journal of Educational Measurement, 30*(3), 253-268.

Rennie, D. (2003). Editorial peer review: Its development and rationale. In F. Godlee & T. Jefferson (Eds.), *Peer review in health sciences* (2nd ed.) (1-13). London: BMJ.

Resnik, D. B., Gutierrez-Ford, C., & Peddada, S. (2008). Perceptions of ethical problems with scientific journal peer review: An exploratory study. *Science and Engineering Ethics, 14*(3), 305-310.

Rothwell, P. M., & Martyn, C. N. (2000). Reproducibility of peer review in clinical neuroscience Is agreement between reviewers any greater than would be expected by chance alone?. *Brain*, *123*(9), 1964-1969.

Rowland, F. (2002). The peer-review process. *Learned Publishing, 15*(4), 247-258.

Rubin, H. R., Redelmeier, D. A., Wu, A. W., & Steinberg, E. P. (1993). How reliable is peer review of scientific abstracts?. *Journal of General Internal Medicine, 8*(5), 255-258.

Schroter, S., Black, N., Evans, S., Carpenter, J., Godlee, F., & Smith, R. (2004). Effects of training on quality of peer review: Randomised controlled trial. *BMJ, 328*(7441), 673.

Siegelman, S. S. (1991). Assassins and zealots: variations in peer review. Special report. *Radiology, 178*(3), 637-642.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: Sage Publications.

Smith, R. (1994). Promoting research into peer review. *BMJ, 309*(6948), 143-144.

Smith, R. (2006). Peer review: A flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine, 99*(4), 178-182.

Snell, L., & Spencer, J. (2005). Reviewers' perceptions of the peer review process for a medical education journal. *Medical Education, 39*(1), 90-97.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, *9*(3), 239-261.

Suls, J., & Martin, R. (2009). The air we breathe: A critical look at practices and alternatives in the peer-review process. *Perspectives on Psychological Science, 4*(1), 40-50.

van Rooyen, S., Black, N., & Godlee, F. (1999). Development of the review quality instrument (RQI) for assessing peer reviews of manuscripts. *Journal of Clinical Epidemiology, 52*(7), 625-629.

van Rooyen, S., Godlee, F., Evans, S., Black, N., & Smith, R. (1999). Effect of open peer review on quality of reviews and on reviewers' recommendations: A randomised trial. *BMJ, 318*(7175), 23-27.

van Rooyen, S., Godlee, F., Evans, S., Smith, R., & Black, N. (1998). Effect of blinding and unmasking on the quality of peer review: a randomized trial. *JAMA, 280*(3), 234-237.

Wang, Z., & Yao, L. (2013). The Effects of Rater Severity and Rater Distribution on Examinees' Ability Estimation for Constructed-Response Items. *ETS Research Report Series*, *2013*(2), i-22.

Ware, M. (2008). Peer review: Benefits, perceptions and alternatives. *Publishing Research Consortium, 4.*

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and Generalizability Theory. *Handbook of Statistics*, *26*, 81-124.

Wilson, H. G. (1988). Parameter estimation for peer grading under incomplete design. *Educational and Psychological Measurement, 48*(1), 69-81.

Wilson, M., & Case, H. (2000). An examination of variation in rater severity over time: A study in rater drift. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice* (Vol. 5). Stamford, CT: Ablex Publishing Corporation.

Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics*, *26*(3), 283-306.

Wind, S. A., Engelhard Jr, G., & Wesolowski, B. (2016). Exploring the effects of rater linking designs and rater fit on achievement estimates within the context of music performance assessments. *Educational Assessment*, *21*(4), 278-299.

Wright, B.D. (1998). Model selection: Rating Scale Model (RSM) or Partial Credit Model (PCM)? *Rasch Measurement Transactions, 12*(3), 641-642.

Wright, B. D., & Tennant, A. (1996). Sample size again. *Rasch Measurement Transactions, 9*(4), 468.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, *46*, 35-51.

Yun, G. J., Donahue, L. M., Dudley, N. M., & McFarland, L. A. (2005). Rater personality, rating format, and social context: Implications for performance appraisal ratings. *International Journal of Selection and Assessment*, *13*(2), 97-107.

# APPENDIX A

## DECISION STUDIES

Table A1. Decision Studies for Varying Numbers of Reviewers and Items

| Reviewers | Items | Expected Observed Score | $\sigma^2_{Rel}$ | $\sigma^2_{Abs}$ | $\rho^2$ | $\Phi$ |
|---|---|---|---|---|---|---|
| 2 | 5 | 0.3329 | 0.2134 | 0.2432 | 0.3590 | 0.3295 |
| 3 | 5 | 0.2641 | 0.1446 | 0.1744 | 0.4526 | 0.4067 |
| 4 | 5 | 0.2297 | 0.1102 | 0.1400 | 0.5204 | 0.4606 |
| 5 | 5 | 0.2090 | 0.0895 | 0.1193 | 0.5718 | 0.5005 |
| 2 | 6 | 0.3263 | 0.2068 | 0.2316 | 0.3663 | 0.3404 |
| 3 | 6 | 0.2593 | 0.1398 | 0.1646 | 0.4610 | 0.4207 |
| 4 | 6 | 0.2258 | 0.1063 | 0.1311 | 0.5293 | 0.4769 |
| 5 | 6 | 0.2057 | 0.0862 | 0.1110 | 0.5811 | 0.5185 |
| 2 | 7 | 0.3216 | 0.2020 | 0.2233 | 0.3717 | 0.3486 |
| 3 | 7 | 0.2559 | 0.1363 | 0.1576 | 0.4671 | 0.4313 |
| 4 | 7 | 0.2230 | 0.1035 | 0.1248 | 0.5359 | 0.4893 |
| 5 | 7 | 0.2033 | 0.0838 | 0.1051 | 0.5879 | 0.5322 |
| 2 | 8 | 0.3180 | 0.1985 | 0.2171 | 0.3758 | 0.3551 |
| 3 | 8 | 0.2533 | 0.1338 | 0.1524 | 0.4719 | 0.4396 |
| 4 | 8 | 0.2209 | 0.1014 | 0.1200 | 0.5410 | 0.4990 |
| 5 | 8 | 0.2015 | 0.0820 | 0.1006 | 0.5931 | 0.5430 |
| 2 | 9 | 0.3153 | 0.1957 | 0.2123 | 0.3791 | 0.3602 |
| 3 | 9 | 0.2513 | 0.1318 | 0.1483 | 0.4756 | 0.4463 |
| 4 | 9 | 0.2193 | 0.0998 | 0.1163 | 0.5450 | 0.5068 |
| 5 | 9 | 0.2001 | 0.0806 | 0.0972 | 0.5972 | 0.5516 |
| 2 | 10 | 0.3130 | 0.1935 | 0.2084 | 0.3818 | 0.3645 |
| 3 | 10 | 0.2497 | 0.1302 | 0.1451 | 0.4787 | 0.4517 |
| 4 | 10 | 0.2180 | 0.0985 | 0.1134 | 0.5482 | 0.5132 |
| 5 | 10 | 0.1990 | 0.0795 | 0.0944 | 0.6006 | 0.5588 |
| 6 | 5 | 0.1953 | 0.0758 | 0.1055 | 0.6121 | 0.5311 |
| 7 | 5 | 0.1854 | 0.0659 | 0.0957 | 0.6445 | 0.5553 |
| 8 | 5 | 0.1781 | 0.0585 | 0.0883 | 0.6712 | 0.5750 |
| 9 | 5 | 0.1723 | 0.0528 | 0.0826 | 0.6936 | 0.5914 |
| 10 | 5 | 0.1677 | 0.0482 | 0.0780 | 0.7125 | 0.6051 |
| 6 | 6 | 0.1923 | 0.0728 | 0.0976 | 0.6216 | 0.5505 |
| 7 | 6 | 0.1827 | 0.0632 | 0.0880 | 0.6541 | 0.5759 |

| Reviewers | Items | Expected Observed Score | $\sigma^2_{Rel}$ | $\sigma^2_{Abs}$ | $\rho^2$ | $\Phi$ |
|---|---|---|---|---|---|---|
| 8 | 6 | 0.1755 | 0.0560 | 0.0808 | 0.6809 | 0.5965 |
| 9 | 6 | 0.1700 | 0.0504 | 0.0753 | 0.7032 | 0.6136 |
| 10 | 6 | 0.1655 | 0.0460 | 0.0708 | 0.7222 | 0.6280 |
| 6 | 7 | 0.1902 | 0.0706 | 0.0919 | 0.6285 | 0.5653 |
| 7 | 7 | 0.1808 | 0.0613 | 0.0825 | 0.6611 | 0.5915 |
| 8 | 7 | 0.1737 | 0.0542 | 0.0755 | 0.6879 | 0.6129 |
| 9 | 7 | 0.1683 | 0.0487 | 0.0700 | 0.7103 | 0.6306 |
| 10 | 7 | 0.1639 | 0.0444 | 0.0656 | 0.7293 | 0.6455 |
| 6 | 8 | 0.1886 | 0.0691 | 0.0877 | 0.6338 | 0.5769 |
| 7 | 8 | 0.1793 | 0.0598 | 0.0784 | 0.6665 | 0.6038 |
| 8 | 8 | 0.1724 | 0.0529 | 0.0715 | 0.6933 | 0.6258 |
| 9 | 8 | 0.1670 | 0.0475 | 0.0661 | 0.7157 | 0.6439 |
| 10 | 8 | 0.1627 | 0.0432 | 0.0618 | 0.7347 | 0.6593 |
| 6 | 9 | 0.1873 | 0.0678 | 0.0844 | 0.6380 | 0.5862 |
| 7 | 9 | 0.1782 | 0.0587 | 0.0752 | 0.6707 | 0.6138 |
| 8 | 9 | 0.1713 | 0.0518 | 0.0684 | 0.6976 | 0.6361 |
| 9 | 9 | 0.1660 | 0.0465 | 0.0630 | 0.7200 | 0.6547 |
| 10 | 9 | 0.1617 | 0.0422 | 0.0588 | 0.7390 | 0.6704 |
| 6 | 10 | 0.1863 | 0.0668 | 0.0817 | 0.6414 | 0.5940 |
| 7 | 10 | 0.1773 | 0.0578 | 0.0727 | 0.6742 | 0.6219 |
| 8 | 10 | 0.1705 | 0.0510 | 0.0659 | 0.7010 | 0.6447 |
| 9 | 10 | 0.1652 | 0.0457 | 0.0606 | 0.7234 | 0.6636 |
| 9 | 10 | 0.1652 | 0.0457 | 0.0606 | 0.7234 | 0.6636 |
| 11 | 5 | 0.1640 | 0.0445 | 0.0743 | 0.7288 | 0.6168 |
| 12 | 5 | 0.1609 | 0.0413 | 0.0711 | 0.7430 | 0.6269 |
| 13 | 5 | 0.1582 | 0.0387 | 0.0685 | 0.7555 | 0.6358 |
| 11 | 6 | 0.1618 | 0.0423 | 0.0671 | 0.7385 | 0.6403 |
| 12 | 6 | 0.1588 | 0.0393 | 0.0641 | 0.7527 | 0.6509 |
| 13 | 6 | 0.1562 | 0.0367 | 0.0615 | 0.7651 | 0.6602 |
| 11 | 7 | 0.1603 | 0.0408 | 0.0621 | 0.7456 | 0.6582 |
| 12 | 7 | 0.1573 | 0.0378 | 0.0591 | 0.7598 | 0.6693 |
| 13 | 7 | 0.1548 | 0.0353 | 0.0565 | 0.7722 | 0.6789 |
| 11 | 8 | 0.1592 | 0.0396 | 0.0582 | 0.7510 | 0.6724 |
| 12 | 8 | 0.1562 | 0.0367 | 0.0553 | 0.7651 | 0.6837 |
| 13 | 8 | 0.1537 | 0.0342 | 0.0528 | 0.7775 | 0.6935 |
| 11 | 9 | 0.1583 | 0.0387 | 0.0553 | 0.7552 | 0.6838 |
| 12 | 9 | 0.1554 | 0.0358 | 0.0524 | 0.7694 | 0.6953 |

| Reviewers | Items | Expected Observed Score | $\sigma^2_{Rel}$ | $\sigma^2_{Abs}$ | $\rho^2$ | $\Phi$ |
|---|---|---|---|---|---|---|
| 13 | 9 | 0.1529 | 0.0334 | 0.0499 | 0.7818 | 0.7054 |
| 11 | 10 | 0.1575 | 0.0380 | 0.0529 | 0.7587 | 0.6932 |
| 12 | 10 | 0.1547 | 0.0351 | 0.0500 | 0.7728 | 0.7049 |
| 13 | 10 | 0.1522 | 0.0327 | 0.0476 | 0.7852 | 0.7152 |
| 14 | 5 | 0.1559 | 0.0364 | 0.0662 | 0.7664 | 0.6435 |
| 15 | 5 | 0.1540 | 0.0345 | 0.0642 | 0.7762 | 0.6504 |
| 16 | 5 | 0.1523 | 0.0327 | 0.0625 | 0.7850 | 0.6566 |
| 14 | 6 | 0.1540 | 0.0345 | 0.0593 | 0.7761 | 0.6684 |
| 15 | 6 | 0.1521 | 0.0326 | 0.0574 | 0.7859 | 0.6756 |
| 16 | 6 | 0.1504 | 0.0309 | 0.0557 | 0.7946 | 0.6821 |
| 14 | 7 | 0.1526 | 0.0331 | 0.0544 | 0.7831 | 0.6873 |
| 15 | 7 | 0.1508 | 0.0312 | 0.0525 | 0.7929 | 0.6948 |
| 16 | 7 | 0.1491 | 0.0296 | 0.0509 | 0.8016 | 0.7015 |
| 14 | 8 | 0.1516 | 0.0321 | 0.0507 | 0.7885 | 0.7022 |
| 15 | 8 | 0.1497 | 0.0302 | 0.0488 | 0.7982 | 0.7100 |
| 16 | 8 | 0.1481 | 0.0286 | 0.0472 | 0.8069 | 0.7168 |
| 14 | 9 | 0.1508 | 0.0313 | 0.0478 | 0.7927 | 0.7143 |
| 15 | 9 | 0.1490 | 0.0294 | 0.0460 | 0.8024 | 0.7222 |
| 16 | 9 | 0.1474 | 0.0278 | 0.0444 | 0.8111 | 0.7292 |
| 14 | 10 | 0.1501 | 0.0306 | 0.0455 | 0.7961 | 0.7243 |
| 15 | 10 | 0.1483 | 0.0288 | 0.0437 | 0.8058 | 0.7323 |
| 16 | 10 | 0.1467 | 0.0272 | 0.0421 | 0.8145 | 0.7395 |
| 17 | 5 | 0.1507 | 0.0312 | 0.0610 | 0.7929 | 0.6621 |
| 18 | 5 | 0.1494 | 0.0299 | 0.0597 | 0.8001 | 0.6671 |
| 19 | 5 | 0.1482 | 0.0287 | 0.0584 | 0.8066 | 0.6716 |
| 20 | 5 | 0.1471 | 0.0276 | 0.0574 | 0.8126 | 0.6757 |
| 17 | 6 | 0.1489 | 0.0294 | 0.0542 | 0.8025 | 0.6879 |
| 18 | 6 | 0.1476 | 0.0281 | 0.0529 | 0.8096 | 0.6931 |
| 19 | 6 | 0.1465 | 0.0269 | 0.0518 | 0.8161 | 0.6979 |
| 20 | 6 | 0.1454 | 0.0259 | 0.0507 | 0.8221 | 0.7022 |
| 17 | 7 | 0.1477 | 0.0281 | 0.0494 | 0.8095 | 0.7075 |
| 18 | 7 | 0.1464 | 0.0269 | 0.0481 | 0.8166 | 0.7130 |
| 19 | 7 | 0.1452 | 0.0257 | 0.0470 | 0.8231 | 0.7179 |
| 20 | 7 | 0.1442 | 0.0247 | 0.0459 | 0.8290 | 0.7224 |
| 17 | 8 | 0.1467 | 0.0272 | 0.0458 | 0.8148 | 0.7230 |
| 18 | 8 | 0.1454 | 0.0259 | 0.0445 | 0.8219 | 0.7286 |
| 19 | 8 | 0.1443 | 0.0248 | 0.0434 | 0.8284 | 0.7337 |

| Reviewers | Items | Expected Observed Score | $\sigma^2_{Rel}$ | $\sigma^2_{Abs}$ | $\rho^2$ | $\Phi$ |
|---|---|---|---|---|---|---|
| 20 | 8 | 0.1433 | 0.0237 | 0.0424 | 0.8343 | 0.7383 |
| 17 | 9 | 0.1459 | 0.0264 | 0.0430 | 0.8190 | 0.7356 |
| 18 | 9 | 0.1447 | 0.0252 | 0.0417 | 0.8261 | 0.7413 |
| 19 | 9 | 0.1436 | 0.0240 | 0.0406 | 0.8325 | 0.7465 |
| 20 | 9 | 0.1426 | 0.0230 | 0.0396 | 0.8384 | 0.7512 |
| 20 | 10 | 0.1420 | 0.0225 | 0.0374 | 0.8418 | 0.7619 |
| 20 | 10 | 0.1420 | 0.0225 | 0.0374 | 0.8418 | 0.7619 |
| 20 | 10 | 0.1420 | 0.0225 | 0.0374 | 0.8418 | 0.7619 |
| 20 | 10 | 0.1420 | 0.0225 | 0.0374 | 0.8418 | 0.7619 |
| 21 | 5 | 0.1461 | 0.0266 | 0.0564 | 0.8180 | 0.6795 |
| 22 | 5 | 0.1452 | 0.0257 | 0.0555 | 0.8231 | 0.6830 |
| 23 | 5 | 0.1444 | 0.0249 | 0.0547 | 0.8277 | 0.6862 |
| 21 | 6 | 0.1444 | 0.0249 | 0.0497 | 0.8275 | 0.7062 |
| 22 | 6 | 0.1436 | 0.0240 | 0.0489 | 0.8325 | 0.7098 |
| 23 | 6 | 0.1428 | 0.0233 | 0.0481 | 0.8372 | 0.7132 |
| 21 | 7 | 0.1432 | 0.0237 | 0.0450 | 0.8344 | 0.7265 |
| 22 | 7 | 0.1424 | 0.0229 | 0.0441 | 0.8394 | 0.7303 |
| 23 | 7 | 0.1416 | 0.0221 | 0.0434 | 0.8440 | 0.7338 |
| 21 | 8 | 0.1423 | 0.0228 | 0.0414 | 0.8397 | 0.7426 |
| 22 | 8 | 0.1415 | 0.0220 | 0.0406 | 0.8447 | 0.7465 |
| 23 | 8 | 0.1407 | 0.0212 | 0.0398 | 0.8493 | 0.7501 |
| 21 | 9 | 0.1416 | 0.0221 | 0.0387 | 0.8438 | 0.7556 |
| 22 | 9 | 0.1408 | 0.0213 | 0.0378 | 0.8488 | 0.7596 |
| 23 | 9 | 0.1401 | 0.0205 | 0.0371 | 0.8534 | 0.7632 |
| 21 | 10 | 0.1411 | 0.0216 | 0.0365 | 0.8472 | 0.7663 |
| 22 | 10 | 0.1403 | 0.0207 | 0.0356 | 0.8521 | 0.7703 |
| 23 | 10 | 0.1395 | 0.0200 | 0.0349 | 0.8567 | 0.7741 |
| 24 | 5 | 0.1437 | 0.0241 | 0.0539 | 0.8320 | 0.6891 |
| 25 | 5 | 0.1430 | 0.0234 | 0.0532 | 0.8360 | 0.6919 |
| 26 | 5 | 0.1423 | 0.0228 | 0.0526 | 0.8398 | 0.6944 |
| 24 | 6 | 0.1420 | 0.0225 | 0.0473 | 0.8415 | 0.7163 |
| 25 | 6 | 0.1414 | 0.0219 | 0.0467 | 0.8454 | 0.7192 |
| 26 | 6 | 0.1408 | 0.0212 | 0.0461 | 0.8492 | 0.7219 |
| 24 | 7 | 0.1409 | 0.0214 | 0.0426 | 0.8483 | 0.7370 |
| 25 | 7 | 0.1402 | 0.0207 | 0.0420 | 0.8523 | 0.7400 |
| 26 | 7 | 0.1396 | 0.0201 | 0.0414 | 0.8560 | 0.7428 |
| 24 | 8 | 0.1400 | 0.0205 | 0.0391 | 0.8536 | 0.7534 |

| Reviewers | Items | Expected Observed Score | $\sigma^2_{Rel}$ | $\sigma^2_{Abs}$ | $\rho^2$ | $\Phi$ |
|---|---|---|---|---|---|---|
| 25 | 8 | 0.1394 | 0.0199 | 0.0385 | 0.8575 | 0.7565 |
| 26 | 8 | 0.1388 | 0.0193 | 0.0379 | 0.8612 | 0.7594 |
| 24 | 9 | 0.1394 | 0.0198 | 0.0364 | 0.8577 | 0.7666 |
| 25 | 9 | 0.1387 | 0.0192 | 0.0357 | 0.8616 | 0.7698 |
| 26 | 9 | 0.1381 | 0.0186 | 0.0352 | 0.8653 | 0.7727 |
| 24 | 10 | 0.1388 | 0.0193 | 0.0342 | 0.8610 | 0.7776 |
| 25 | 10 | 0.1382 | 0.0187 | 0.0336 | 0.8649 | 0.7808 |
| 26 | 10 | 0.1376 | 0.0181 | 0.0330 | 0.8686 | 0.7838 |
| 27 | 5 | 0.1417 | 0.0222 | 0.0520 | 0.8432 | 0.6968 |
| 28 | 5 | 0.1412 | 0.0217 | 0.0515 | 0.8465 | 0.6990 |
| 29 | 5 | 0.1407 | 0.0212 | 0.0510 | 0.8496 | 0.7011 |
| 27 | 6 | 0.1402 | 0.0207 | 0.0455 | 0.8526 | 0.7244 |
| 28 | 6 | 0.1397 | 0.0201 | 0.0450 | 0.8559 | 0.7267 |
| 29 | 6 | 0.1392 | 0.0196 | 0.0445 | 0.8589 | 0.7289 |
| 27 | 7 | 0.1391 | 0.0196 | 0.0408 | 0.8595 | 0.7454 |
| 28 | 7 | 0.1385 | 0.0190 | 0.0403 | 0.8627 | 0.7479 |
| 29 | 7 | 0.1381 | 0.0185 | 0.0398 | 0.8657 | 0.7501 |
| 27 | 8 | 0.1382 | 0.0187 | 0.0373 | 0.8647 | 0.7620 |
| 28 | 8 | 0.1377 | 0.0182 | 0.0368 | 0.8679 | 0.7645 |
| 29 | 8 | 0.1372 | 0.0177 | 0.0363 | 0.8709 | 0.7669 |
| 27 | 9 | 0.1376 | 0.0181 | 0.0346 | 0.8687 | 0.7755 |
| 28 | 9 | 0.1371 | 0.0176 | 0.0341 | 0.8720 | 0.7780 |
| 29 | 9 | 0.1366 | 0.0171 | 0.0336 | 0.8750 | 0.7804 |
| 27 | 10 | 0.1371 | 0.0175 | 0.0324 | 0.8720 | 0.7866 |
| 28 | 10 | 0.1366 | 0.0170 | 0.0319 | 0.8752 | 0.7892 |
| 29 | 10 | 0.1361 | 0.0166 | 0.0315 | 0.8783 | 0.7916 |
| 30 | 5 | 0.1402 | 0.0207 | 0.0505 | 0.8524 | 0.7031 |
| 31 | 5 | 0.1398 | 0.0203 | 0.0500 | 0.8552 | 0.7049 |
| 32 | 5 | 0.1394 | 0.0198 | 0.0496 | 0.8577 | 0.7067 |
| 30 | 6 | 0.1387 | 0.0192 | 0.0440 | 0.8618 | 0.7310 |
| 31 | 6 | 0.1383 | 0.0187 | 0.0436 | 0.8645 | 0.7329 |
| 32 | 6 | 0.1379 | 0.0183 | 0.0432 | 0.8670 | 0.7347 |
| 30 | 7 | 0.1376 | 0.0181 | 0.0394 | 0.8686 | 0.7523 |
| 31 | 7 | 0.1372 | 0.0177 | 0.0389 | 0.8713 | 0.7543 |
| 32 | 7 | 0.1368 | 0.0173 | 0.0385 | 0.8738 | 0.7562 |
| 30 | 8 | 0.1368 | 0.0173 | 0.0359 | 0.8737 | 0.7691 |
| 31 | 8 | 0.1364 | 0.0169 | 0.0355 | 0.8764 | 0.7712 |

| Reviewers | Items | Expected Observed Score | $\sigma^2_{Rel}$ | $\sigma^2_{Abs}$ | $\rho^2$ | $\Phi$ |
|---|---|---|---|---|---|---|
| 32 | 8 | 0.1360 | 0.0165 | 0.0351 | 0.8789 | 0.7731 |
| 30 | 9 | 0.1362 | 0.0166 | 0.0332 | 0.8778 | 0.7827 |
| 31 | 9 | 0.1358 | 0.0162 | 0.0328 | 0.8805 | 0.7848 |
| 32 | 9 | 0.1354 | 0.0158 | 0.0324 | 0.8830 | 0.7868 |
| 30 | 10 | 0.1357 | 0.0161 | 0.0310 | 0.8811 | 0.7939 |
| 31 | 10 | 0.1352 | 0.0157 | 0.0306 | 0.8837 | 0.7961 |
| 32 | 10 | 0.1349 | 0.0153 | 0.0302 | 0.8863 | 0.7981 |
| 33 | 5 | 0.1390 | 0.0194 | 0.0492 | 0.8601 | 0.7083 |
| 33 | 6 | 0.1375 | 0.0180 | 0.0428 | 0.8694 | 0.7365 |
| 33 | 7 | 0.1364 | 0.0169 | 0.0382 | 0.8762 | 0.7580 |
| 33 | 8 | 0.1356 | 0.0161 | 0.0347 | 0.8813 | 0.7750 |
| 33 | 9 | 0.1350 | 0.0155 | 0.0320 | 0.8854 | 0.7887 |
| 33 | 10 | 0.1345 | 0.0150 | 0.0299 | 0.8886 | 0.8000 |
| 10 | 10 | 0.1610 | 0.0415 | 0.0564 | 0.7424 | 0.6795 |
| 17 | 10 | 0.1453 | 0.0258 | 0.0407 | 0.8223 | 0.7459 |
| 18 | 10 | 0.1441 | 0.0246 | 0.0395 | 0.8294 | 0.7517 |
| 19 | 10 | 0.1430 | 0.0235 | 0.0384 | 0.8359 | 0.7570 |

APPENDIX B

DECISION STUDIES WITH ITEM AS A FIXED FACET

Table A2. Decision Studies for Five Items and Varying Numbers of Reviewers

| Reviewers | Items | Expected Observed Score | $\sigma^2_{Rel}$ | $\sigma^2_{Abs}$ | $\rho^2$ | $\Phi$ |
|---|---|---|---|---|---|---|
| 2 | 5 | 0.1195 | 0.2932 | 0.1736 | 0.1736 | 0.0010 |
| 3 | 5 | 0.1195 | 0.2353 | 0.1158 | 0.1158 | 0.0008 |
| 4 | 5 | 0.1195 | 0.2063 | 0.0868 | 0.0868 | 0.0007 |
| 5 | 5 | 0.1195 | 0.1890 | 0.0695 | 0.0695 | 0.0006 |
| 6 | 5 | 0.1195 | 0.1774 | 0.0579 | 0.0579 | 0.0006 |
| 7 | 5 | 0.1195 | 0.1691 | 0.0496 | 0.0496 | 0.0006 |
| 8 | 5 | 0.1195 | 0.1629 | 0.0434 | 0.0434 | 0.0005 |
| 9 | 5 | 0.1195 | 0.1581 | 0.0386 | 0.0386 | 0.0005 |
| 10 | 5 | 0.1195 | 0.1543 | 0.0347 | 0.0347 | 0.0005 |
| 11 | 5 | 0.1195 | 0.1511 | 0.0316 | 0.0316 | 0.0005 |
| 12 | 5 | 0.1195 | 0.1485 | 0.0289 | 0.0289 | 0.0005 |
| 13 | 5 | 0.1195 | 0.1462 | 0.0267 | 0.0267 | 0.0005 |
| 14 | 5 | 0.1195 | 0.1443 | 0.0248 | 0.0248 | 0.0005 |
| 15 | 5 | 0.1195 | 0.1427 | 0.0232 | 0.0232 | 0.0005 |
| 16 | 5 | 0.1195 | 0.1412 | 0.0217 | 0.0217 | 0.0005 |
| 17 | 5 | 0.1195 | 0.1400 | 0.0204 | 0.0204 | 0.0005 |
| 18 | 5 | 0.1195 | 0.1388 | 0.0193 | 0.0193 | 0.0005 |
| 19 | 5 | 0.1195 | 0.1378 | 0.0183 | 0.0183 | 0.0005 |
| 20 | 5 | 0.1195 | 0.1369 | 0.0174 | 0.0174 | 0.0005 |
| 21 | 5 | 0.1195 | 0.1361 | 0.0165 | 0.0165 | 0.0005 |
| 22 | 5 | 0.1195 | 0.1353 | 0.0158 | 0.0158 | 0.0005 |
| 23 | 5 | 0.1195 | 0.1346 | 0.0151 | 0.0151 | 0.0005 |
| 24 | 5 | 0.1195 | 0.1340 | 0.0145 | 0.0145 | 0.0005 |
| 25 | 5 | 0.1195 | 0.1334 | 0.0139 | 0.0139 | 0.0004 |
| 26 | 5 | 0.1195 | 0.1329 | 0.0134 | 0.0134 | 0.0004 |
| 27 | 5 | 0.1195 | 0.1324 | 0.0129 | 0.0129 | 0.0004 |
| 28 | 5 | 0.1195 | 0.1319 | 0.0124 | 0.0124 | 0.0004 |
| 29 | 5 | 0.1195 | 0.1315 | 0.0120 | 0.0120 | 0.0004 |
| 30 | 5 | 0.1195 | 0.1311 | 0.0116 | 0.0116 | 0.0004 |

APPENDIX C

PREDICTED DECISION CATEGORIES USING AVERAGE RAW TOTAL

Table A3. Observed and Predicted Frequencies Using Average Raw Total Score

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Average Raw Total | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| 6.00 | Accept/Minor Revision | 0 | 0.00 | -0.03 | 0.00 | 0.00 |
| | Major Revision | 0 | 0.09 | -0.31 | 0.00 | 0.09 |
| | Reject | 1 | 0.91 | 0.31 | 1.00 | 0.91 |
| 7.00 | Accept/Minor Revision | 0 | 0.00 | -0.07 | 0.00 | 0.00 |
| | Major Revision | 0 | 0.36 | -0.64 | 0.00 | 0.12 |
| | Reject | 3 | 2.63 | 0.65 | 1.00 | 0.88 |
| 8.00 | Accept/Minor Revision | 0 | 0.00 | -0.06 | 0.00 | 0.00 |
| | Major Revision | 0 | 0.16 | -0.44 | 0.00 | 0.16 |
| | Reject | 1 | 0.84 | 0.44 | 1.00 | 0.84 |
| 8.50 | Accept/Minor Revision | 0 | 0.01 | -0.09 | 0.00 | 0.00 |
| | Major Revision | 1 | 0.37 | 1.15 | 0.50 | 0.18 |
| | Reject | 1 | 1.62 | -1.12 | 0.50 | 0.81 |
| 9.00 | Accept/Minor Revision | 0 | 0.01 | -0.11 | 0.00 | 0.01 |
| | Major Revision | 0 | 0.42 | -0.73 | 0.00 | 0.21 |
| | Reject | 2 | 1.56 | 0.75 | 1.00 | 0.78 |
| 9.33 | Accept/Minor Revision | 0 | 0.01 | -0.09 | 0.00 | 0.01 |
| | Major Revision | 0 | 0.23 | -0.55 | 0.00 | 0.23 |
| | Reject | 1 | 0.76 | 0.56 | 1.00 | 0.76 |
| 10.00 | Accept/Minor Revision | 0 | 0.09 | -0.30 | 0.00 | 0.01 |
| | Major Revision | 2 | 2.17 | -0.14 | 0.25 | 0.27 |
| | Reject | 6 | 5.73 | 0.21 | 0.75 | 0.72 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Average Raw Total | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| 10.50 | Accept/Minor Revision | 0 | 0.05 | -0.22 | 0.00 | 0.02 |
| | Major Revision | 1 | 0.92 | 0.11 | 0.33 | 0.31 |
| | Reject | 2 | 2.04 | -0.05 | 0.67 | 0.68 |
| 10.67 | Accept/Minor Revision | 0 | 0.03 | -0.19 | 0.00 | 0.02 |
| | Major Revision | 1 | 0.63 | 0.56 | 0.50 | 0.32 |
| | Reject | 1 | 1.33 | -0.50 | 0.50 | 0.67 |
| 11.00 | Accept/Minor Revision | 0 | 0.14 | -0.38 | 0.00 | 0.02 |
| | Major Revision | 3 | 2.38 | 0.49 | 0.43 | 0.34 |
| | Reject | 4 | 4.48 | -0.37 | 0.57 | 0.64 |
| 11.33 | Accept/Minor Revision | 0 | 0.03 | -0.16 | 0.00 | 0.03 |
| | Major Revision | 0 | 0.36 | -0.76 | 0.00 | 0.36 |
| | Reject | 1 | 0.61 | 0.80 | 1.00 | 0.61 |
| 11.50 | Accept/Minor Revision | 0 | 0.11 | -0.34 | 0.00 | 0.03 |
| | Major Revision | 0 | 1.50 | -1.55 | 0.00 | 0.38 |
| | Reject | 4 | 2.39 | 1.64 | 1.00 | 0.60 |
| 11.67 | Accept/Minor Revision | 0 | 0.06 | -0.25 | 0.00 | 0.03 |
| | Major Revision | 1 | 0.78 | 0.33 | 0.50 | 0.39 |
| | Reject | 1 | 1.16 | -0.23 | 0.50 | 0.58 |
| 12.00 | Accept/Minor Revision | 0 | 0.88 | -0.95 | 0.00 | 0.04 |
| | Major Revision | 12 | 9.88 | 0.88 | 0.50 | 0.41 |
| | Reject | 12 | 13.24 | -0.51 | 0.50 | 0.55 |
| 12.25 | Accept/Minor Revision | 0 | 0.04 | -0.21 | 0.00 | 0.04 |
| | Major Revision | 0 | 0.43 | -0.87 | 0.00 | 0.43 |
| | Reject | 1 | 0.53 | 0.94 | 1.00 | 0.53 |
| 12.50 | Accept/Minor Revision | 0 | 0.33 | -0.59 | 0.00 | 0.05 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Average Raw Total | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| | Major Revision | 5 | 3.13 | 1.42 | 0.71 | 0.45 |
| | Reject | 2 | 3.54 | -1.16 | 0.29 | 0.51 |
| 12.67 | Accept/Minor Revision | 0 | 0.05 | -0.23 | 0.00 | 0.05 |
| | Major Revision | 0 | 0.46 | -0.92 | 0.00 | 0.46 |
| | Reject | 1 | 0.49 | 1.02 | 1.00 | 0.49 |
| 13.00 | Accept/Minor Revision | 0 | 1.12 | -1.09 | 0.00 | 0.06 |
| | Major Revision | 10 | 8.63 | 0.64 | 0.56 | 0.48 |
| | Reject | 8 | 8.25 | -0.12 | 0.44 | 0.46 |
| 13.33 | Accept/Minor Revision | 0 | 0.22 | -0.49 | 0.00 | 0.07 |
| | Major Revision | 1 | 1.50 | -0.58 | 0.33 | 0.50 |
| | Reject | 2 | 1.28 | 0.84 | 0.67 | 0.43 |
| 13.50 | Accept/Minor Revision | 1 | 0.64 | 0.47 | 0.13 | 0.08 |
| | Major Revision | 4 | 4.08 | -0.05 | 0.50 | 0.51 |
| | Reject | 3 | 3.29 | -0.21 | 0.38 | 0.41 |
| 13.67 | Accept/Minor Revision | 0 | 0.35 | -0.61 | 0.00 | 0.09 |
| | Major Revision | 1 | 2.07 | -1.07 | 0.25 | 0.52 |
| | Reject | 3 | 1.58 | 1.45 | 0.75 | 0.40 |
| 14.00 | Accept/Minor Revision | 3 | 2.02 | 0.73 | 0.15 | 0.10 |
| | Major Revision | 10 | 10.70 | -0.31 | 0.50 | 0.53 |
| | Reject | 7 | 7.28 | -0.13 | 0.35 | 0.36 |
| 14.33 | Accept/Minor Revision | 0 | 0.12 | -0.36 | 0.00 | 0.12 |
| | Major Revision | 1 | 0.55 | 0.91 | 1.00 | 0.55 |
| | Reject | 0 | 0.33 | -0.71 | 0.00 | 0.33 |
| 14.50 | Accept/Minor Revision | 2 | 0.89 | 1.27 | 0.29 | 0.13 |
| | Major Revision | 4 | 3.88 | 0.09 | 0.57 | 0.55 |
| | Reject | 1 | 2.23 | -1.00 | 0.14 | 0.32 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Average Raw Total | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| 14.67 | Accept/Minor Revision | 0 | 0.95 | -1.05 | 0.00 | 0.14 |
| | Major Revision | 6 | 3.92 | 1.58 | 0.86 | 0.56 |
| | Reject | 1 | 2.13 | -0.93 | 0.14 | 0.30 |
| 14.75 | Accept/Minor Revision | 0 | 0.14 | -0.41 | 0.00 | 0.14 |
| | Major Revision | 1 | 0.56 | 0.88 | 1.00 | 0.56 |
| | Reject | 0 | 0.30 | -0.65 | 0.00 | 0.30 |
| 15.00 | Accept/Minor Revision | 6 | 3.60 | 1.38 | 0.26 | 0.16 |
| | Major Revision | 12 | 13.06 | -0.45 | 0.52 | 0.57 |
| | Reject | 5 | 6.34 | -0.63 | 0.22 | 0.28 |
| 15.33 | Accept/Minor Revision | 0 | 0.36 | -0.66 | 0.00 | 0.18 |
| | Major Revision | 1 | 1.15 | -0.21 | 0.50 | 0.57 |
| | Reject | 1 | 0.50 | 0.82 | 0.50 | 0.25 |
| 15.50 | Accept/Minor Revision | 2 | 0.96 | 1.19 | 0.40 | 0.19 |
| | Major Revision | 3 | 2.87 | 0.12 | 0.60 | 0.57 |
| | Reject | 0 | 1.18 | -1.24 | 0.00 | 0.24 |
| 15.67 | Accept/Minor Revision | 2 | 1.63 | 0.33 | 0.25 | 0.20 |
| | Major Revision | 4 | 4.59 | -0.42 | 0.50 | 0.57 |
| | Reject | 2 | 1.78 | 0.19 | 0.25 | 0.22 |
| 15.75 | Accept/Minor Revision | 1 | 0.42 | 1.01 | 0.50 | 0.21 |
| | Major Revision | 1 | 1.15 | -0.21 | 0.50 | 0.57 |
| | Reject | 0 | 0.43 | -0.74 | 0.00 | 0.22 |
| 16.00 | Accept/Minor Revision | 3 | 4.37 | -0.75 | 0.16 | 0.23 |
| | Major Revision | 9 | 10.87 | -0.87 | 0.47 | 0.57 |
| | Reject | 7 | 3.76 | 1.87 | 0.37 | 0.20 |
| 16.33 | Accept/Minor Revision | 1 | 0.52 | 0.78 | 0.50 | 0.26 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Average Raw Total | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| | Major Revision | 1 | 1.13 | -0.19 | 0.50 | 0.57 |
| | Reject | 0 | 0.35 | -0.65 | 0.00 | 0.18 |
| 16.50 | Accept/Minor Revision | 0 | 2.46 | -1.84 | 0.00 | 0.27 |
| | Major Revision | 7 | 5.06 | 1.30 | 0.78 | 0.56 |
| | Reject | 2 | 1.48 | 0.47 | 0.22 | 0.16 |
| 16.67 | Accept/Minor Revision | 1 | 0.58 | 0.66 | 0.50 | 0.29 |
| | Major Revision | 1 | 1.12 | -0.16 | 0.50 | 0.56 |
| | Reject | 0 | 0.31 | -0.60 | 0.00 | 0.15 |
| 16.75 | Accept/Minor Revision | 0 | 0.30 | -0.65 | 0.00 | 0.30 |
| | Major Revision | 1 | 0.55 | 0.90 | 1.00 | 0.55 |
| | Reject | 0 | 0.15 | -0.42 | 0.00 | 0.15 |
| 17.00 | Accept/Minor Revision | 6 | 6.40 | -0.19 | 0.30 | 0.32 |
| | Major Revision | 11 | 10.91 | 0.04 | 0.55 | 0.55 |
| | Reject | 3 | 2.69 | 0.20 | 0.15 | 0.13 |
| 17.33 | Accept/Minor Revision | 1 | 1.06 | -0.07 | 0.33 | 0.35 |
| | Major Revision | 2 | 1.59 | 0.47 | 0.67 | 0.53 |
| | Reject | 0 | 0.35 | -0.63 | 0.00 | 0.12 |
| 17.50 | Accept/Minor Revision | 3 | 2.96 | 0.03 | 0.38 | 0.37 |
| | Major Revision | 4 | 4.17 | -0.12 | 0.50 | 0.52 |
| | Reject | 1 | 0.87 | 0.15 | 0.13 | 0.11 |
| 17.67 | Accept/Minor Revision | 0 | 0.39 | -0.79 | 0.00 | 0.39 |
| | Major Revision | 1 | 0.51 | 0.98 | 1.00 | 0.51 |
| | Reject | 0 | 0.10 | -0.33 | 0.00 | 0.10 |
| 17.75 | Accept/Minor Revision | 0 | 0.40 | -0.81 | 0.00 | 0.40 |
| | Major Revision | 1 | 0.51 | 0.99 | 1.00 | 0.51 |
| | Reject | 0 | 0.10 | -0.33 | 0.00 | 0.10 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Average Raw Total | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| 18.00 | Accept/Minor Revision | 12 | 7.59 | 2.11 | 0.67 | 0.42 |
| | Major Revision | 4 | 8.86 | -2.29 | 0.22 | 0.49 |
| | Reject | 2 | 1.56 | 0.37 | 0.11 | 0.09 |
| 18.25 | Accept/Minor Revision | 0 | 0.45 | -0.90 | 0.00 | 0.45 |
| | Major Revision | 1 | 0.48 | 1.05 | 1.00 | 0.48 |
| | Reject | 0 | 0.08 | -0.29 | 0.00 | 0.08 |
| 18.33 | Accept/Minor Revision | 1 | 0.46 | 1.09 | 1.00 | 0.46 |
| | Major Revision | 0 | 0.47 | -0.94 | 0.00 | 0.47 |
| | Reject | 0 | 0.07 | -0.28 | 0.00 | 0.07 |
| 18.50 | Accept/Minor Revision | 3 | 1.90 | 1.11 | 0.75 | 0.47 |
| | Major Revision | 1 | 1.83 | -0.84 | 0.25 | 0.46 |
| | Reject | 0 | 0.27 | -0.54 | 0.00 | 0.07 |
| 18.67 | Accept/Minor Revision | 0 | 0.49 | -0.98 | 0.00 | 0.49 |
| | Major Revision | 1 | 0.45 | 1.11 | 1.00 | 0.45 |
| | Reject | 0 | 0.06 | -0.26 | 0.00 | 0.06 |
| 18.75 | Accept/Minor Revision | 0 | 0.50 | -1.00 | 0.00 | 0.50 |
| | Major Revision | 1 | 0.44 | 1.13 | 1.00 | 0.44 |
| | Reject | 0 | 0.06 | -0.25 | 0.00 | 0.06 |
| 19.00 | Accept/Minor Revision | 5 | 3.68 | 1.00 | 0.71 | 0.53 |
| | Major Revision | 0 | 2.95 | -2.26 | 0.00 | 0.42 |
| | Reject | 2 | 0.37 | 2.76 | 0.29 | 0.05 |
| 19.25 | Accept/Minor Revision | 0 | 0.55 | -1.11 | 0.00 | 0.55 |
| | Major Revision | 1 | 0.40 | 1.22 | 1.00 | 0.40 |
| | Reject | 0 | 0.05 | -0.22 | 0.00 | 0.05 |
| 19.33 | Accept/Minor Revision | 0 | 0.56 | -1.13 | 0.00 | 0.56 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Average Raw Total | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| | Major Revision | 1 | 0.40 | 1.24 | 1.00 | 0.40 |
| | Reject | 0 | 0.04 | -0.22 | 0.00 | 0.04 |
| 19.40 | Accept/Minor Revision | 0 | 0.57 | -1.14 | 0.00 | 0.57 |
| | Major Revision | 1 | 0.39 | 1.25 | 1.00 | 0.39 |
| | Reject | 0 | 0.04 | -0.21 | 0.00 | 0.04 |
| 19.50 | Accept/Minor Revision | 1 | 1.15 | -0.22 | 0.50 | 0.58 |
| | Major Revision | 1 | 0.76 | 0.34 | 0.50 | 0.38 |
| | Reject | 0 | 0.08 | -0.29 | 0.00 | 0.04 |
| 19.67 | Accept/Minor Revision | 1 | 0.59 | 0.83 | 1.00 | 0.59 |
| | Major Revision | 0 | 0.37 | -0.77 | 0.00 | 0.37 |
| | Reject | 0 | 0.04 | -0.20 | 0.00 | 0.04 |
| 20.00 | Accept/Minor Revision | 3 | 3.76 | -0.64 | 0.50 | 0.63 |
| | Major Revision | 3 | 2.06 | 0.81 | 0.50 | 0.34 |
| | Reject | 0 | 0.18 | -0.44 | 0.00 | 0.03 |
| 20.50 | Accept/Minor Revision | 1 | 1.34 | -0.52 | 0.50 | 0.67 |
| | Major Revision | 1 | 0.61 | 0.60 | 0.50 | 0.30 |
| | Reject | 0 | 0.05 | -0.22 | 0.00 | 0.02 |
| 20.67 | Accept/Minor Revision | 1 | 0.69 | 0.68 | 1.00 | 0.69 |
| | Major Revision | 0 | 0.29 | -0.64 | 0.00 | 0.29 |
| | Reject | 0 | 0.02 | -0.15 | 0.00 | 0.02 |
| 20.75 | Accept/Minor Revision | 0 | 0.69 | -1.51 | 0.00 | 0.69 |
| | Major Revision | 1 | 0.29 | 1.58 | 1.00 | 0.29 |
| | Reject | 0 | 0.02 | -0.14 | 0.00 | 0.02 |
| 21.00 | Accept/Minor Revision | 4 | 2.86 | 1.26 | 1.00 | 0.71 |
| | Major Revision | 0 | 1.07 | -1.21 | 0.00 | 0.27 |
| | Reject | 0 | 0.07 | -0.26 | 0.00 | 0.02 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Average Raw Total | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| 22.00 | Accept/Minor Revision | 1 | 1.58 | -1.00 | 0.50 | 0.79 |
| | Major Revision | 1 | 0.41 | 1.05 | 0.50 | 0.20 |
| | Reject | 0 | 0.02 | -0.14 | 0.00 | 0.01 |
| 25.00 | Accept/Minor Revision | 0 | 0.92 | -3.45 | 0.00 | 0.92 |
| | Major Revision | 1 | 0.08 | 3.48 | 1.00 | 0.08 |
| | Reject | 0 | 0.00 | -0.04 | 0.00 | 0.00 |

APPENDIX D

PREDICTED DECISION CATEGORIES USING PUBLISHABILITY MEASURE

Table A4. Observed and Predicted Frequencies Using Manuscript Publishability Measure

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Publishability Measure | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| -4.88 | Accept/Minor Revision | 0 | 0.01 | -0.07 | 0.00 | 0.01 |
| | Major Revision | 0 | 0.11 | -0.35 | 0.00 | 0.11 |
| | Reject | 1 | 0.89 | 0.36 | 1.00 | 0.89 |
| -4.82 | Accept/Minor Revision | 0 | 0.01 | -0.08 | 0.00 | 0.01 |
| | Major Revision | 0 | 0.11 | -0.35 | 0.00 | 0.11 |
| | Reject | 1 | 0.88 | 0.36 | 1.00 | 0.88 |
| -3.96 | Accept/Minor Revision | 0 | 0.01 | -0.11 | 0.00 | 0.01 |
| | Major Revision | 0 | 0.16 | -0.44 | 0.00 | 0.16 |
| | Reject | 1 | 0.82 | 0.46 | 1.00 | 0.82 |
| -3.89 | Accept/Minor Revision | 0 | 0.03 | -0.16 | 0.00 | 0.01 |
| | Major Revision | 1 | 0.34 | 1.24 | 0.50 | 0.17 |
| | Reject | 1 | 1.64 | -1.16 | 0.50 | 0.82 |
| -3.72 | Accept/Minor Revision | 0 | 0.01 | -0.12 | 0.00 | 0.01 |
| | Major Revision | 0 | 0.18 | -0.47 | 0.00 | 0.18 |
| | Reject | 1 | 0.80 | 0.50 | 1.00 | 0.80 |
| -3.68 | Accept/Minor Revision | 0 | 0.01 | -0.12 | 0.00 | 0.01 |
| | Major Revision | 0 | 0.19 | -0.48 | 0.00 | 0.19 |
| | Reject | 1 | 0.80 | 0.50 | 1.00 | 0.80 |
| -3.58 | Accept/Minor Revision | 0 | 0.02 | -0.13 | 0.00 | 0.02 |
| | Major Revision | 0 | 0.19 | -0.49 | 0.00 | 0.19 |
| | Reject | 1 | 0.79 | 0.52 | 1.00 | 0.79 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Publishability Measure | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| -3.33 | Accept/Minor Revision | 0 | 0.02 | -0.14 | 0.00 | 0.02 |
| | Major Revision | 0 | 0.22 | -0.52 | 0.00 | 0.22 |
| | Reject | 1 | 0.76 | 0.55 | 1.00 | 0.76 |
| -3.23 | Accept/Minor Revision | 0 | 0.02 | -0.15 | 0.00 | 0.02 |
| | Major Revision | 0 | 0.22 | -0.54 | 0.00 | 0.22 |
| | Reject | 1 | 0.75 | 0.57 | 1.00 | 0.75 |
| -3.10 | Accept/Minor Revision | 0 | 0.02 | -0.16 | 0.00 | 0.02 |
| | Major Revision | 0 | 0.24 | -0.56 | 0.00 | 0.24 |
| | Reject | 1 | 0.74 | 0.59 | 1.00 | 0.74 |
| -2.95 | Accept/Minor Revision | 0 | 0.03 | -0.17 | 0.00 | 0.03 |
| | Major Revision | 0 | 0.25 | -0.58 | 0.00 | 0.25 |
| | Reject | 1 | 0.72 | 0.62 | 1.00 | 0.72 |
| -2.80 | Accept/Minor Revision | 0 | 0.03 | -0.18 | 0.00 | 0.03 |
| | Major Revision | 0 | 0.27 | -0.60 | 0.00 | 0.27 |
| | Reject | 1 | 0.70 | 0.65 | 1.00 | 0.70 |
| -2.62 | Accept/Minor Revision | 0 | 0.03 | -0.19 | 0.00 | 0.03 |
| | Major Revision | 0 | 0.28 | -0.63 | 0.00 | 0.28 |
| | Reject | 1 | 0.68 | 0.68 | 1.00 | 0.68 |
| -2.52 | Accept/Minor Revision | 0 | 0.04 | -0.20 | 0.00 | 0.04 |
| | Major Revision | 0 | 0.29 | -0.64 | 0.00 | 0.29 |
| | Reject | 1 | 0.67 | 0.70 | 1.00 | 0.67 |
| -2.44 | Accept/Minor Revision | 0 | 0.04 | -0.20 | 0.00 | 0.04 |
| | Major Revision | 1 | 0.30 | 1.52 | 1.00 | 0.30 |
| | Reject | 0 | 0.66 | -1.39 | 0.00 | 0.66 |
| -2.40 | Accept/Minor Revision | 0 | 0.04 | -0.21 | 0.00 | 0.04 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Publishability Measure | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| | Major Revision | 1 | 0.31 | 1.51 | 1.00 | 0.31 |
| | Reject | 0 | 0.65 | -1.37 | 0.00 | 0.65 |
| -2.31 | Accept/Minor Revision | 0 | 0.04 | -0.21 | 0.00 | 0.04 |
| | Major Revision | 1 | 0.32 | 1.47 | 1.00 | 0.32 |
| | Reject | 0 | 0.64 | -1.34 | 0.00 | 0.64 |
| -2.25 | Accept/Minor Revision | 0 | 0.05 | -0.22 | 0.00 | 0.05 |
| | Major Revision | 0 | 0.32 | -0.69 | 0.00 | 0.32 |
| | Reject | 1 | 0.63 | 0.76 | 1.00 | 0.63 |
| -2.24 | Accept/Minor Revision | 0 | 0.05 | -0.22 | 0.00 | 0.05 |
| | Major Revision | 0 | 0.32 | -0.69 | 0.00 | 0.32 |
| | Reject | 1 | 0.63 | 0.76 | 1.00 | 0.63 |
| -2.13 | Accept/Minor Revision | 0 | 0.05 | -0.23 | 0.00 | 0.05 |
| | Major Revision | 0 | 0.33 | -0.71 | 0.00 | 0.33 |
| | Reject | 1 | 0.62 | 0.79 | 1.00 | 0.62 |
| -2.10 | Accept/Minor Revision | 0 | 0.05 | -0.23 | 0.00 | 0.05 |
| | Major Revision | 0 | 0.34 | -0.71 | 0.00 | 0.34 |
| | Reject | 1 | 0.61 | 0.80 | 1.00 | 0.61 |
| -2.00 | Accept/Minor Revision | 0 | 0.05 | -0.24 | 0.00 | 0.05 |
| | Major Revision | 1 | 0.35 | 1.37 | 1.00 | 0.35 |
| | Reject | 0 | 0.60 | -1.22 | 0.00 | 0.60 |
| -1.87 | Accept/Minor Revision | 0 | 0.06 | -0.25 | 0.00 | 0.06 |
| | Major Revision | 1 | 0.36 | 1.33 | 1.00 | 0.36 |
| | Reject | 0 | 0.58 | -1.17 | 0.00 | 0.58 |
| -1.85 | Accept/Minor Revision | 1 | 0.24 | 1.59 | 0.25 | 0.06 |
| | Major Revision | 1 | 1.45 | -0.47 | 0.25 | 0.36 |
| | Reject | 2 | 2.30 | -0.31 | 0.50 | 0.58 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Publishability Measure | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| -1.84 | Accept/Minor Revision | 1 | 0.06 | 3.93 | 1.00 | 0.06 |
| | Major Revision | 0 | 0.36 | -0.76 | 0.00 | 0.36 |
| | Reject | 0 | 0.57 | -1.16 | 0.00 | 0.57 |
| -1.82 | Accept/Minor Revision | 0 | 0.06 | -0.26 | 0.00 | 0.06 |
| | Major Revision | 0 | 0.37 | -0.76 | 0.00 | 0.37 |
| | Reject | 1 | 0.57 | 0.87 | 1.00 | 0.57 |
| -1.73 | Accept/Minor Revision | 0 | 0.07 | -0.26 | 0.00 | 0.07 |
| | Major Revision | 1 | 0.38 | 1.29 | 1.00 | 0.38 |
| | Reject | 0 | 0.56 | -1.12 | 0.00 | 0.56 |
| -1.72 | Accept/Minor Revision | 0 | 0.07 | -0.27 | 0.00 | 0.07 |
| | Major Revision | 1 | 0.38 | 1.28 | 1.00 | 0.38 |
| | Reject | 0 | 0.56 | -1.12 | 0.00 | 0.56 |
| -1.66 | Accept/Minor Revision | 0 | 0.14 | -0.38 | 0.00 | 0.07 |
| | Major Revision | 0 | 0.77 | -1.12 | 0.00 | 0.38 |
| | Reject | 2 | 1.10 | 1.28 | 1.00 | 0.55 |
| -1.65 | Accept/Minor Revision | 0 | 0.14 | -0.39 | 0.00 | 0.07 |
| | Major Revision | 0 | 0.77 | -1.12 | 0.00 | 0.38 |
| | Reject | 2 | 1.09 | 1.29 | 1.00 | 0.55 |
| -1.61 | Accept/Minor Revision | 0 | 0.07 | -0.28 | 0.00 | 0.07 |
| | Major Revision | 1 | 0.39 | 1.25 | 1.00 | 0.39 |
| | Reject | 0 | 0.54 | -1.08 | 0.00 | 0.54 |
| -1.60 | Accept/Minor Revision | 0 | 0.07 | -0.28 | 0.00 | 0.07 |
| | Major Revision | 0 | 0.39 | -0.80 | 0.00 | 0.39 |
| | Reject | 1 | 0.54 | 0.93 | 1.00 | 0.54 |
| -1.57 | Accept/Minor Revision | 0 | 0.07 | -0.28 | 0.00 | 0.07 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Publishability Measure | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| | Major Revision | 1 | 0.39 | 1.24 | 1.00 | 0.39 |
| | Reject | 0 | 0.53 | -1.07 | 0.00 | 0.53 |
| -1.45 | Accept/Minor Revision | 0 | 0.08 | -0.29 | 0.00 | 0.08 |
| | Major Revision | 1 | 0.41 | 1.21 | 1.00 | 0.41 |
| | Reject | 0 | 0.52 | -1.03 | 0.00 | 0.52 |
| -1.39 | Accept/Minor Revision | 0 | 0.08 | -0.30 | 0.00 | 0.08 |
| | Major Revision | 0 | 0.41 | -0.84 | 0.00 | 0.41 |
| | Reject | 1 | 0.51 | 0.99 | 1.00 | 0.51 |
| -1.38 | Accept/Minor Revision | 0 | 0.08 | -0.30 | 0.00 | 0.08 |
| | Major Revision | 0 | 0.41 | -0.84 | 0.00 | 0.41 |
| | Reject | 1 | 0.51 | 0.99 | 1.00 | 0.51 |
| -1.35 | Accept/Minor Revision | 0 | 0.34 | -0.61 | 0.00 | 0.08 |
| | Major Revision | 1 | 1.66 | -0.67 | 0.25 | 0.42 |
| | Reject | 3 | 2.00 | 1.00 | 0.75 | 0.50 |
| -1.31 | Accept/Minor Revision | 0 | 0.09 | -0.31 | 0.00 | 0.09 |
| | Major Revision | 1 | 0.42 | 1.18 | 1.00 | 0.42 |
| | Reject | 0 | 0.49 | -0.99 | 0.00 | 0.49 |
| -1.27 | Accept/Minor Revision | 0 | 0.18 | -0.44 | 0.00 | 0.09 |
| | Major Revision | 0 | 0.85 | -1.21 | 0.00 | 0.42 |
| | Reject | 2 | 0.98 | 1.45 | 1.00 | 0.49 |
| -1.26 | Accept/Minor Revision | 0 | 0.09 | -0.31 | 0.00 | 0.09 |
| | Major Revision | 1 | 0.42 | 1.17 | 1.00 | 0.42 |
| | Reject | 0 | 0.49 | -0.97 | 0.00 | 0.49 |
| -1.25 | Accept/Minor Revision | 0 | 0.18 | -0.44 | 0.00 | 0.09 |
| | Major Revision | 2 | 0.85 | 1.64 | 1.00 | 0.43 |
| | Reject | 0 | 0.97 | -1.37 | 0.00 | 0.49 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Publishability Measure | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| -1.23 | Accept/Minor Revision | 0 | 0.09 | -0.32 | 0.00 | 0.09 |
| | Major Revision | 0 | 0.43 | -0.86 | 0.00 | 0.43 |
| | Reject | 1 | 0.48 | 1.04 | 1.00 | 0.48 |
| -1.18 | Accept/Minor Revision | 0 | 0.19 | -0.45 | 0.00 | 0.09 |
| | Major Revision | 1 | 0.86 | 0.19 | 0.50 | 0.43 |
| | Reject | 1 | 0.95 | 0.07 | 0.50 | 0.47 |
| -1.17 | Accept/Minor Revision | 0 | 0.09 | -0.32 | 0.00 | 0.09 |
| | Major Revision | 1 | 0.43 | 1.14 | 1.00 | 0.43 |
| | Reject | 0 | 0.47 | -0.95 | 0.00 | 0.47 |
| -1.11 | Accept/Minor Revision | 0 | 0.10 | -0.33 | 0.00 | 0.10 |
| | Major Revision | 1 | 0.44 | 1.13 | 1.00 | 0.44 |
| | Reject | 0 | 0.46 | -0.93 | 0.00 | 0.46 |
| -1.08 | Accept/Minor Revision | 0 | 0.10 | -0.33 | 0.00 | 0.10 |
| | Major Revision | 0 | 0.44 | -0.89 | 0.00 | 0.44 |
| | Reject | 1 | 0.46 | 1.08 | 1.00 | 0.46 |
| -1.05 | Accept/Minor Revision | 0 | 0.10 | -0.34 | 0.00 | 0.10 |
| | Major Revision | 0 | 0.44 | -0.89 | 0.00 | 0.44 |
| | Reject | 1 | 0.45 | 1.09 | 1.00 | 0.45 |
| -1.04 | Accept/Minor Revision | 0 | 0.20 | -0.48 | 0.00 | 0.10 |
| | Major Revision | 2 | 0.89 | 1.58 | 1.00 | 0.44 |
| | Reject | 0 | 0.91 | -1.29 | 0.00 | 0.45 |
| -1.02 | Accept/Minor Revision | 0 | 0.10 | -0.34 | 0.00 | 0.10 |
| | Major Revision | 1 | 0.45 | 1.11 | 1.00 | 0.45 |
| | Reject | 0 | 0.45 | -0.90 | 0.00 | 0.45 |
| -0.99 | Accept/Minor Revision | 0 | 0.10 | -0.34 | 0.00 | 0.10 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Publishability Measure | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| | Major Revision | 0 | 0.45 | -0.90 | 0.00 | 0.45 |
| | Reject | 1 | 0.45 | 1.12 | 1.00 | 0.45 |
| -0.98 | Accept/Minor Revision | 0 | 0.11 | -0.34 | 0.00 | 0.11 |
| | Major Revision | 0 | 0.45 | -0.91 | 0.00 | 0.45 |
| | Reject | 1 | 0.44 | 1.12 | 1.00 | 0.44 |
| -0.94 | Accept/Minor Revision | 0 | 0.11 | -0.35 | 0.00 | 0.11 |
| | Major Revision | 0 | 0.45 | -0.91 | 0.00 | 0.45 |
| | Reject | 1 | 0.44 | 1.13 | 1.00 | 0.44 |
| -0.93 | Accept/Minor Revision | 0 | 0.11 | -0.35 | 0.00 | 0.11 |
| | Major Revision | 1 | 0.45 | 1.09 | 1.00 | 0.45 |
| | Reject | 0 | 0.44 | -0.88 | 0.00 | 0.44 |
| -0.87 | Accept/Minor Revision | 0 | 1.24 | -1.18 | 0.00 | 0.11 |
| | Major Revision | 4 | 5.06 | -0.64 | 0.36 | 0.46 |
| | Reject | 7 | 4.70 | 1.40 | 0.64 | 0.43 |
| -0.86 | Accept/Minor Revision | 1 | 0.23 | 1.73 | 0.50 | 0.11 |
| | Major Revision | 1 | 0.92 | 0.11 | 0.50 | 0.46 |
| | Reject | 0 | 0.85 | -1.22 | 0.00 | 0.43 |
| -0.85 | Accept/Minor Revision | 0 | 0.11 | -0.36 | 0.00 | 0.11 |
| | Major Revision | 1 | 0.46 | 1.08 | 1.00 | 0.46 |
| | Reject | 0 | 0.42 | -0.86 | 0.00 | 0.42 |
| -0.84 | Accept/Minor Revision | 0 | 0.11 | -0.36 | 0.00 | 0.11 |
| | Major Revision | 1 | 0.46 | 1.08 | 1.00 | 0.46 |
| | Reject | 0 | 0.42 | -0.86 | 0.00 | 0.42 |
| -0.83 | Accept/Minor Revision | 0 | 0.12 | -0.36 | 0.00 | 0.12 |
| | Major Revision | 1 | 0.46 | 1.08 | 1.00 | 0.46 |
| | Reject | 0 | 0.42 | -0.85 | 0.00 | 0.42 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Publishability Measure | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| -0.82 | Accept/Minor Revision | 0 | 0.12 | -0.36 | 0.00 | 0.12 |
| | Major Revision | 1 | 0.46 | 1.07 | 1.00 | 0.46 |
| | Reject | 0 | 0.42 | -0.85 | 0.00 | 0.42 |
| -0.81 | Accept/Minor Revision | 0 | 0.12 | -0.36 | 0.00 | 0.12 |
| | Major Revision | 1 | 0.47 | 1.07 | 1.00 | 0.47 |
| | Reject | 0 | 0.42 | -0.85 | 0.00 | 0.42 |
| -0.77 | Accept/Minor Revision | 0 | 0.36 | -0.64 | 0.00 | 0.12 |
| | Major Revision | 2 | 1.41 | 0.69 | 0.67 | 0.47 |
| | Reject | 1 | 1.24 | -0.28 | 0.33 | 0.41 |
| -0.76 | Accept/Minor Revision | 0 | 0.12 | -0.37 | 0.00 | 0.12 |
| | Major Revision | 1 | 0.47 | 1.06 | 1.00 | 0.47 |
| | Reject | 0 | 0.41 | -0.83 | 0.00 | 0.41 |
| -0.74 | Accept/Minor Revision | 0 | 0.12 | -0.37 | 0.00 | 0.12 |
| | Major Revision | 1 | 0.47 | 1.06 | 1.00 | 0.47 |
| | Reject | 0 | 0.41 | -0.83 | 0.00 | 0.41 |
| -0.72 | Accept/Minor Revision | 0 | 0.12 | -0.37 | 0.00 | 0.12 |
| | Major Revision | 1 | 0.47 | 1.06 | 1.00 | 0.47 |
| | Reject | 0 | 0.40 | -0.82 | 0.00 | 0.40 |
| -0.69 | Accept/Minor Revision | 0 | 0.12 | -0.38 | 0.00 | 0.12 |
| | Major Revision | 1 | 0.48 | 1.05 | 1.00 | 0.48 |
| | Reject | 0 | 0.40 | -0.82 | 0.00 | 0.40 |
| -0.63 | Accept/Minor Revision | 1 | 0.26 | 1.56 | 0.50 | 0.13 |
| | Major Revision | 0 | 0.96 | -1.36 | 0.00 | 0.48 |
| | Reject | 1 | 0.78 | 0.32 | 0.50 | 0.39 |
| -0.59 | Accept/Minor Revision | 1 | 0.13 | 2.56 | 1.00 | 0.13 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Publishability Measure | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| | Major Revision | 0 | 0.48 | -0.97 | 0.00 | 0.48 |
| | Reject | 0 | 0.38 | -0.79 | 0.00 | 0.38 |
| -0.58 | Accept/Minor Revision | 0 | 0.13 | -0.39 | 0.00 | 0.13 |
| | Major Revision | 0 | 0.48 | -0.97 | 0.00 | 0.48 |
| | Reject | 1 | 0.38 | 1.27 | 1.00 | 0.38 |
| -0.57 | Accept/Minor Revision | 0 | 0.13 | -0.39 | 0.00 | 0.13 |
| | Major Revision | 0 | 0.48 | -0.97 | 0.00 | 0.48 |
| | Reject | 1 | 0.38 | 1.27 | 1.00 | 0.38 |
| -0.56 | Accept/Minor Revision | 0 | 0.13 | -0.39 | 0.00 | 0.13 |
| | Major Revision | 0 | 0.49 | -0.97 | 0.00 | 0.49 |
| | Reject | 1 | 0.38 | 1.28 | 1.00 | 0.38 |
| -0.55 | Accept/Minor Revision | 1 | 0.27 | 1.51 | 0.50 | 0.14 |
| | Major Revision | 1 | 0.97 | 0.04 | 0.50 | 0.49 |
| | Reject | 0 | 0.76 | -1.10 | 0.00 | 0.38 |
| -0.54 | Accept/Minor Revision | 1 | 0.14 | 2.52 | 1.00 | 0.14 |
| | Major Revision | 0 | 0.49 | -0.97 | 0.00 | 0.49 |
| | Reject | 0 | 0.38 | -0.78 | 0.00 | 0.38 |
| -0.53 | Accept/Minor Revision | 0 | 0.14 | -0.40 | 0.00 | 0.14 |
| | Major Revision | 1 | 0.49 | 1.03 | 1.00 | 0.49 |
| | Reject | 0 | 0.38 | -0.78 | 0.00 | 0.38 |
| -0.52 | Accept/Minor Revision | 0 | 0.55 | -0.80 | 0.00 | 0.14 |
| | Major Revision | 2 | 1.95 | 0.05 | 0.50 | 0.49 |
| | Reject | 2 | 1.50 | 0.52 | 0.50 | 0.37 |
| -0.50 | Accept/Minor Revision | 0 | 0.14 | -0.40 | 0.00 | 0.14 |
| | Major Revision | 0 | 0.49 | -0.98 | 0.00 | 0.49 |
| | Reject | 1 | 0.37 | 1.30 | 1.00 | 0.37 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Publishability Measure | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| -0.49 | Accept/Minor Revision | 0 | 0.14 | -0.40 | 0.00 | 0.14 |
| | Major Revision | 0 | 0.49 | -0.98 | 0.00 | 0.49 |
| | Reject | 1 | 0.37 | 1.30 | 1.00 | 0.37 |
| -0.48 | Accept/Minor Revision | 0 | 0.14 | -0.40 | 0.00 | 0.14 |
| | Major Revision | 1 | 0.49 | 1.02 | 1.00 | 0.49 |
| | Reject | 0 | 0.37 | -0.76 | 0.00 | 0.37 |
| -0.44 | Accept/Minor Revision | 0 | 0.14 | -0.41 | 0.00 | 0.14 |
| | Major Revision | 1 | 0.49 | 1.01 | 1.00 | 0.49 |
| | Reject | 0 | 0.36 | -0.75 | 0.00 | 0.36 |
| -0.43 | Accept/Minor Revision | 0 | 0.14 | -0.41 | 0.00 | 0.14 |
| | Major Revision | 1 | 0.49 | 1.01 | 1.00 | 0.49 |
| | Reject | 0 | 0.36 | -0.75 | 0.00 | 0.36 |
| -0.41 | Accept/Minor Revision | 1 | 0.15 | 2.42 | 1.00 | 0.15 |
| | Major Revision | 0 | 0.50 | -0.99 | 0.00 | 0.50 |
| | Reject | 0 | 0.36 | -0.75 | 0.00 | 0.36 |
| -0.39 | Accept/Minor Revision | 0 | 0.59 | -0.83 | 0.00 | 0.15 |
| | Major Revision | 1 | 1.99 | -0.99 | 0.25 | 0.50 |
| | Reject | 3 | 1.42 | 1.65 | 0.75 | 0.36 |
| -0.35 | Accept/Minor Revision | 0 | 0.15 | -0.42 | 0.00 | 0.15 |
| | Major Revision | 0 | 0.50 | -1.00 | 0.00 | 0.50 |
| | Reject | 1 | 0.35 | 1.36 | 1.00 | 0.35 |
| -0.32 | Accept/Minor Revision | 0 | 0.31 | -0.60 | 0.00 | 0.15 |
| | Major Revision | 1 | 1.00 | -0.01 | 0.50 | 0.50 |
| | Reject | 1 | 0.69 | 0.46 | 0.50 | 0.35 |
| -0.31 | Accept/Minor Revision | 0 | 0.15 | -0.43 | 0.00 | 0.15 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Publishability Measure | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| | Major Revision | 1 | 0.50 | 1.00 | 1.00 | 0.50 |
| | Reject | 0 | 0.34 | -0.72 | 0.00 | 0.34 |
| -0.29 | Accept/Minor Revision | 1 | 0.16 | 2.33 | 1.00 | 0.16 |
| | Major Revision | 0 | 0.50 | -1.01 | 0.00 | 0.50 |
| | Reject | 0 | 0.34 | -0.72 | 0.00 | 0.34 |
| -0.28 | Accept/Minor Revision | 0 | 0.16 | -0.43 | 0.00 | 0.16 |
| | Major Revision | 1 | 0.50 | 0.99 | 1.00 | 0.50 |
| | Reject | 0 | 0.34 | -0.72 | 0.00 | 0.34 |
| -0.27 | Accept/Minor Revision | 0 | 0.16 | -0.43 | 0.00 | 0.16 |
| | Major Revision | 1 | 0.50 | 0.99 | 1.00 | 0.50 |
| | Reject | 0 | 0.34 | -0.71 | 0.00 | 0.34 |
| -0.26 | Accept/Minor Revision | 0 | 0.32 | -0.61 | 0.00 | 0.16 |
| | Major Revision | 2 | 1.01 | 1.40 | 1.00 | 0.51 |
| | Reject | 0 | 0.67 | -1.01 | 0.00 | 0.34 |
| -0.25 | Accept/Minor Revision | 0 | 0.32 | -0.61 | 0.00 | 0.16 |
| | Major Revision | 1 | 1.01 | -0.02 | 0.50 | 0.51 |
| | Reject | 1 | 0.67 | 0.49 | 0.50 | 0.34 |
| -0.24 | Accept/Minor Revision | 0 | 0.16 | -0.44 | 0.00 | 0.16 |
| | Major Revision | 0 | 0.51 | -1.01 | 0.00 | 0.51 |
| | Reject | 1 | 0.33 | 1.41 | 1.00 | 0.33 |
| -0.23 | Accept/Minor Revision | 1 | 0.16 | 2.29 | 1.00 | 0.16 |
| | Major Revision | 0 | 0.51 | -1.01 | 0.00 | 0.51 |
| | Reject | 0 | 0.33 | -0.71 | 0.00 | 0.33 |
| -0.20 | Accept/Minor Revision | 0 | 0.16 | -0.44 | 0.00 | 0.16 |
| | Major Revision | 0 | 0.51 | -1.02 | 0.00 | 0.51 |
| | Reject | 1 | 0.33 | 1.43 | 1.00 | 0.33 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Publishability Measure | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| -0.18 | Accept/Minor Revision | 0 | 0.16 | -0.44 | 0.00 | 0.16 |
| | Major Revision | 1 | 0.51 | 0.98 | 1.00 | 0.51 |
| | Reject | 0 | 0.33 | -0.69 | 0.00 | 0.33 |
| -0.17 | Accept/Minor Revision | 0 | 0.33 | -0.63 | 0.00 | 0.17 |
| | Major Revision | 1 | 1.02 | -0.03 | 0.50 | 0.51 |
| | Reject | 1 | 0.65 | 0.53 | 0.50 | 0.32 |
| -0.16 | Accept/Minor Revision | 0 | 0.17 | -0.45 | 0.00 | 0.17 |
| | Major Revision | 0 | 0.51 | -1.02 | 0.00 | 0.51 |
| | Reject | 1 | 0.32 | 1.45 | 1.00 | 0.32 |
| -0.14 | Accept/Minor Revision | 0 | 0.17 | -0.45 | 0.00 | 0.17 |
| | Major Revision | 1 | 0.51 | 0.98 | 1.00 | 0.51 |
| | Reject | 0 | 0.32 | -0.69 | 0.00 | 0.32 |
| -0.11 | Accept/Minor Revision | 1 | 0.51 | 0.75 | 0.33 | 0.17 |
| | Major Revision | 2 | 1.54 | 0.53 | 0.67 | 0.51 |
| | Reject | 0 | 0.95 | -1.18 | 0.00 | 0.32 |
| -0.09 | Accept/Minor Revision | 0 | 0.17 | -0.46 | 0.00 | 0.17 |
| | Major Revision | 1 | 0.52 | 0.97 | 1.00 | 0.52 |
| | Reject | 0 | 0.31 | -0.67 | 0.00 | 0.31 |
| -0.07 | Accept/Minor Revision | 1 | 0.35 | 1.22 | 0.50 | 0.17 |
| | Major Revision | 1 | 1.03 | -0.05 | 0.50 | 0.52 |
| | Reject | 0 | 0.62 | -0.95 | 0.00 | 0.31 |
| -0.05 | Accept/Minor Revision | 0 | 0.18 | -0.46 | 0.00 | 0.18 |
| | Major Revision | 1 | 0.52 | 0.97 | 1.00 | 0.52 |
| | Reject | 0 | 0.31 | -0.67 | 0.00 | 0.31 |
| 0.01 | Accept/Minor Revision | 1 | 0.18 | 2.13 | 1.00 | 0.18 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Publishability Measure | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| | Major Revision | 0 | 0.52 | -1.04 | 0.00 | 0.52 |
| | Reject | 0 | 0.30 | -0.65 | 0.00 | 0.30 |
| 0.04 | Accept/Minor Revision | 0 | 0.37 | -0.67 | 0.00 | 0.18 |
| | Major Revision | 1 | 1.04 | -0.06 | 0.50 | 0.52 |
| | Reject | 1 | 0.59 | 0.64 | 0.50 | 0.30 |
| 0.05 | Accept/Minor Revision | 0 | 0.18 | -0.48 | 0.00 | 0.18 |
| | Major Revision | 1 | 0.52 | 0.96 | 1.00 | 0.52 |
| | Reject | 0 | 0.29 | -0.64 | 0.00 | 0.29 |
| 0.06 | Accept/Minor Revision | 0 | 0.19 | -0.48 | 0.00 | 0.19 |
| | Major Revision | 0 | 0.52 | -1.05 | 0.00 | 0.52 |
| | Reject | 1 | 0.29 | 1.56 | 1.00 | 0.29 |
| 0.08 | Accept/Minor Revision | 0 | 0.75 | -0.96 | 0.00 | 0.19 |
| | Major Revision | 2 | 2.09 | -0.09 | 0.50 | 0.52 |
| | Reject | 2 | 1.16 | 0.93 | 0.50 | 0.29 |
| 0.09 | Accept/Minor Revision | 1 | 0.19 | 2.08 | 1.00 | 0.19 |
| | Major Revision | 0 | 0.52 | -1.05 | 0.00 | 0.52 |
| | Reject | 0 | 0.29 | -0.64 | 0.00 | 0.29 |
| 0.12 | Accept/Minor Revision | 0 | 0.19 | -0.49 | 0.00 | 0.19 |
| | Major Revision | 1 | 0.53 | 0.95 | 1.00 | 0.53 |
| | Reject | 0 | 0.28 | -0.63 | 0.00 | 0.28 |
| 0.13 | Accept/Minor Revision | 0 | 0.19 | -0.49 | 0.00 | 0.19 |
| | Major Revision | 1 | 0.53 | 0.95 | 1.00 | 0.53 |
| | Reject | 0 | 0.28 | -0.63 | 0.00 | 0.28 |
| 0.15 | Accept/Minor Revision | 0 | 0.19 | -0.49 | 0.00 | 0.19 |
| | Major Revision | 1 | 0.53 | 0.95 | 1.00 | 0.53 |
| | Reject | 0 | 0.28 | -0.62 | 0.00 | 0.28 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Publishability Measure | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| 0.17 | Accept/Minor Revision | 0 | 0.19 | -0.49 | 0.00 | 0.19 |
| | Major Revision | 1 | 0.53 | 0.95 | 1.00 | 0.53 |
| | Reject | 0 | 0.28 | -0.62 | 0.00 | 0.28 |
| 0.19 | Accept/Minor Revision | 1 | 0.59 | 0.59 | 0.33 | 0.20 |
| | Major Revision | 1 | 1.58 | -0.68 | 0.33 | 0.53 |
| | Reject | 1 | 0.83 | 0.23 | 0.33 | 0.28 |
| 0.22 | Accept/Minor Revision | 0 | 0.20 | -0.50 | 0.00 | 0.20 |
| | Major Revision | 1 | 0.53 | 0.94 | 1.00 | 0.53 |
| | Reject | 0 | 0.27 | -0.61 | 0.00 | 0.27 |
| 0.24 | Accept/Minor Revision | 0 | 0.40 | -0.71 | 0.00 | 0.20 |
| | Major Revision | 1 | 1.06 | -0.08 | 0.50 | 0.53 |
| | Reject | 1 | 0.54 | 0.74 | 0.50 | 0.27 |
| 0.28 | Accept/Minor Revision | 0 | 0.21 | -0.51 | 0.00 | 0.21 |
| | Major Revision | 0 | 0.53 | -1.06 | 0.00 | 0.53 |
| | Reject | 1 | 0.26 | 1.67 | 1.00 | 0.26 |
| 0.29 | Accept/Minor Revision | 0 | 0.21 | -0.51 | 0.00 | 0.21 |
| | Major Revision | 0 | 0.53 | -1.07 | 0.00 | 0.53 |
| | Reject | 1 | 0.26 | 1.68 | 1.00 | 0.26 |
| 0.30 | Accept/Minor Revision | 0 | 0.21 | -0.51 | 0.00 | 0.21 |
| | Major Revision | 0 | 0.53 | -1.07 | 0.00 | 0.53 |
| | Reject | 1 | 0.26 | 1.68 | 1.00 | 0.26 |
| 0.32 | Accept/Minor Revision | 2 | 0.42 | 2.75 | 1.00 | 0.21 |
| | Major Revision | 0 | 1.06 | -1.51 | 0.00 | 0.53 |
| | Reject | 0 | 0.52 | -0.84 | 0.00 | 0.26 |
| 0.34 | Accept/Minor Revision | 0 | 0.21 | -0.52 | 0.00 | 0.21 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Publishability Measure | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| | Major Revision | 0 | 0.53 | -1.07 | 0.00 | 0.53 |
| | Reject | 1 | 0.26 | 1.70 | 1.00 | 0.26 |
| 0.35 | Accept/Minor Revision | 0 | 0.21 | -0.52 | 0.00 | 0.21 |
| | Major Revision | 1 | 0.53 | 0.94 | 1.00 | 0.53 |
| | Reject | 0 | 0.25 | -0.58 | 0.00 | 0.25 |
| 0.36 | Accept/Minor Revision | 0 | 0.21 | -0.52 | 0.00 | 0.21 |
| | Major Revision | 1 | 0.53 | 0.93 | 1.00 | 0.53 |
| | Reject | 0 | 0.25 | -0.58 | 0.00 | 0.25 |
| 0.37 | Accept/Minor Revision | 0 | 0.21 | -0.52 | 0.00 | 0.21 |
| | Major Revision | 1 | 0.53 | 0.93 | 1.00 | 0.53 |
| | Reject | 0 | 0.25 | -0.58 | 0.00 | 0.25 |
| 0.38 | Accept/Minor Revision | 0 | 0.43 | -0.74 | 0.00 | 0.21 |
| | Major Revision | 1 | 1.07 | -0.10 | 0.50 | 0.53 |
| | Reject | 1 | 0.50 | 0.81 | 0.50 | 0.25 |
| 0.40 | Accept/Minor Revision | 0 | 0.22 | -0.53 | 0.00 | 0.22 |
| | Major Revision | 1 | 0.53 | 0.93 | 1.00 | 0.53 |
| | Reject | 0 | 0.25 | -0.58 | 0.00 | 0.25 |
| 0.41 | Accept/Minor Revision | 0 | 0.22 | -0.53 | 0.00 | 0.22 |
| | Major Revision | 1 | 0.54 | 0.93 | 1.00 | 0.54 |
| | Reject | 0 | 0.25 | -0.57 | 0.00 | 0.25 |
| 0.42 | Accept/Minor Revision | 0 | 0.22 | -0.53 | 0.00 | 0.22 |
| | Major Revision | 1 | 0.54 | 0.93 | 1.00 | 0.54 |
| | Reject | 0 | 0.25 | -0.57 | 0.00 | 0.25 |
| 0.43 | Accept/Minor Revision | 0 | 0.22 | -0.53 | 0.00 | 0.22 |
| | Major Revision | 1 | 0.54 | 0.93 | 1.00 | 0.54 |
| | Reject | 0 | 0.24 | -0.57 | 0.00 | 0.24 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Publishability Measure | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| 0.46 | Accept/Minor Revision | 0 | 0.22 | -0.53 | 0.00 | 0.22 |
| | Major Revision | 1 | 0.54 | 0.93 | 1.00 | 0.54 |
| | Reject | 0 | 0.24 | -0.56 | 0.00 | 0.24 |
| 0.47 | Accept/Minor Revision | 0 | 0.22 | -0.54 | 0.00 | 0.22 |
| | Major Revision | 1 | 0.54 | 0.93 | 1.00 | 0.54 |
| | Reject | 0 | 0.24 | -0.56 | 0.00 | 0.24 |
| 0.49 | Accept/Minor Revision | 0 | 0.23 | -0.54 | 0.00 | 0.23 |
| | Major Revision | 1 | 0.54 | 0.93 | 1.00 | 0.54 |
| | Reject | 0 | 0.24 | -0.56 | 0.00 | 0.24 |
| 0.50 | Accept/Minor Revision | 0 | 0.45 | -0.77 | 0.00 | 0.23 |
| | Major Revision | 0 | 1.07 | -1.52 | 0.00 | 0.54 |
| | Reject | 2 | 0.47 | 2.54 | 1.00 | 0.24 |
| 0.51 | Accept/Minor Revision | 0 | 0.23 | -0.54 | 0.00 | 0.23 |
| | Major Revision | 1 | 0.54 | 0.93 | 1.00 | 0.54 |
| | Reject | 0 | 0.24 | -0.55 | 0.00 | 0.24 |
| 0.53 | Accept/Minor Revision | 0 | 0.23 | -0.55 | 0.00 | 0.23 |
| | Major Revision | 1 | 0.54 | 0.93 | 1.00 | 0.54 |
| | Reject | 0 | 0.23 | -0.55 | 0.00 | 0.23 |
| 0.54 | Accept/Minor Revision | 0 | 0.23 | -0.55 | 0.00 | 0.23 |
| | Major Revision | 1 | 0.54 | 0.93 | 1.00 | 0.54 |
| | Reject | 0 | 0.23 | -0.55 | 0.00 | 0.23 |
| 0.55 | Accept/Minor Revision | 0 | 0.93 | -1.10 | 0.00 | 0.23 |
| | Major Revision | 4 | 2.15 | 1.85 | 1.00 | 0.54 |
| | Reject | 0 | 0.92 | -1.09 | 0.00 | 0.23 |
| 0.56 | Accept/Minor Revision | 1 | 0.46 | 0.90 | 0.50 | 0.23 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Publishability | Measure | Observed | Predicted | Pearson Residual | Observed | Predicted |
| | Major Revision | 0 | 1.08 | -1.53 | 0.00 | 0.54 |
| | Reject | 1 | 0.46 | 0.91 | 0.50 | 0.23 |
| 0.57 | Accept/Minor Revision | 1 | 0.23 | 1.81 | 1.00 | 0.23 |
| | Major Revision | 0 | 0.54 | -1.08 | 0.00 | 0.54 |
| | Reject | 0 | 0.23 | -0.54 | 0.00 | 0.23 |
| 0.59 | Accept/Minor Revision | 1 | 0.47 | 0.88 | 0.50 | 0.24 |
| | Major Revision | 0 | 1.08 | -1.53 | 0.00 | 0.54 |
| | Reject | 1 | 0.45 | 0.93 | 0.50 | 0.23 |
| 0.61 | Accept/Minor Revision | 0 | 0.24 | -0.56 | 0.00 | 0.24 |
| | Major Revision | 1 | 0.54 | 0.92 | 1.00 | 0.54 |
| | Reject | 0 | 0.22 | -0.54 | 0.00 | 0.22 |
| 0.62 | Accept/Minor Revision | 1 | 0.48 | 0.87 | 0.50 | 0.24 |
| | Major Revision | 1 | 1.08 | -0.11 | 0.50 | 0.54 |
| | Reject | 0 | 0.44 | -0.76 | 0.00 | 0.22 |
| 0.66 | Accept/Minor Revision | 1 | 0.24 | 1.77 | 1.00 | 0.24 |
| | Major Revision | 0 | 0.54 | -1.08 | 0.00 | 0.54 |
| | Reject | 0 | 0.22 | -0.53 | 0.00 | 0.22 |
| 0.68 | Accept/Minor Revision | 0 | 0.24 | -0.57 | 0.00 | 0.24 |
| | Major Revision | 1 | 0.54 | 0.92 | 1.00 | 0.54 |
| | Reject | 0 | 0.22 | -0.52 | 0.00 | 0.22 |
| 0.72 | Accept/Minor Revision | 0 | 0.25 | -0.58 | 0.00 | 0.25 |
| | Major Revision | 1 | 0.54 | 0.92 | 1.00 | 0.54 |
| | Reject | 0 | 0.21 | -0.52 | 0.00 | 0.21 |
| 0.75 | Accept/Minor Revision | 1 | 0.25 | 1.72 | 1.00 | 0.25 |
| | Major Revision | 0 | 0.54 | -1.08 | 0.00 | 0.54 |
| | Reject | 0 | 0.21 | -0.51 | 0.00 | 0.21 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Publishability Measure | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| 0.77 | Accept/Minor Revision | 0 | 0.25 | -0.58 | 0.00 | 0.25 |
| | Major Revision | 1 | 0.54 | 0.92 | 1.00 | 0.54 |
| | Reject | 0 | 0.21 | -0.51 | 0.00 | 0.21 |
| 0.78 | Accept/Minor Revision | 2 | 0.76 | 1.64 | 0.67 | 0.25 |
| | Major Revision | 0 | 1.62 | -1.88 | 0.00 | 0.54 |
| | Reject | 1 | 0.61 | 0.55 | 0.33 | 0.20 |
| 0.79 | Accept/Minor Revision | 1 | 0.26 | 1.71 | 1.00 | 0.26 |
| | Major Revision | 0 | 0.54 | -1.09 | 0.00 | 0.54 |
| | Reject | 0 | 0.20 | -0.51 | 0.00 | 0.20 |
| 0.80 | Accept/Minor Revision | 3 | 0.77 | 2.95 | 1.00 | 0.26 |
| | Major Revision | 0 | 1.62 | -1.88 | 0.00 | 0.54 |
| | Reject | 0 | 0.61 | -0.87 | 0.00 | 0.20 |
| 0.81 | Accept/Minor Revision | 0 | 0.26 | -0.59 | 0.00 | 0.26 |
| | Major Revision | 0 | 0.54 | -1.09 | 0.00 | 0.54 |
| | Reject | 1 | 0.20 | 1.99 | 1.00 | 0.20 |
| 0.83 | Accept/Minor Revision | 1 | 0.78 | 0.29 | 0.33 | 0.26 |
| | Major Revision | 0 | 1.62 | -1.88 | 0.00 | 0.54 |
| | Reject | 2 | 0.60 | 2.03 | 0.67 | 0.20 |
| 0.84 | Accept/Minor Revision | 1 | 0.52 | 0.77 | 0.50 | 0.26 |
| | Major Revision | 1 | 1.08 | -0.12 | 0.50 | 0.54 |
| | Reject | 0 | 0.40 | -0.70 | 0.00 | 0.20 |
| 0.85 | Accept/Minor Revision | 1 | 0.26 | 1.68 | 1.00 | 0.26 |
| | Major Revision | 0 | 0.54 | -1.09 | 0.00 | 0.54 |
| | Reject | 0 | 0.20 | -0.49 | 0.00 | 0.20 |
| 0.95 | Accept/Minor Revision | 1 | 0.55 | 0.72 | 0.50 | 0.27 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Publishability Measure | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| | Major Revision | 1 | 1.08 | -0.12 | 0.50 | 0.54 |
| | Reject | 0 | 0.37 | -0.68 | 0.00 | 0.19 |
| 0.98 | Accept/Minor Revision | 0 | 0.28 | -0.62 | 0.00 | 0.28 |
| | Major Revision | 1 | 0.54 | 0.92 | 1.00 | 0.54 |
| | Reject | 0 | 0.18 | -0.47 | 0.00 | 0.18 |
| 1.00 | Accept/Minor Revision | 0 | 0.28 | -0.62 | 0.00 | 0.28 |
| | Major Revision | 1 | 0.54 | 0.92 | 1.00 | 0.54 |
| | Reject | 0 | 0.18 | -0.47 | 0.00 | 0.18 |
| 1.02 | Accept/Minor Revision | 0 | 0.28 | -0.62 | 0.00 | 0.28 |
| | Major Revision | 1 | 0.54 | 0.92 | 1.00 | 0.54 |
| | Reject | 0 | 0.18 | -0.47 | 0.00 | 0.18 |
| 1.03 | Accept/Minor Revision | 0 | 1.12 | -1.25 | 0.00 | 0.28 |
| | Major Revision | 4 | 2.16 | 1.84 | 1.00 | 0.54 |
| | Reject | 0 | 0.71 | -0.93 | 0.00 | 0.18 |
| 1.04 | Accept/Minor Revision | 0 | 0.28 | -0.63 | 0.00 | 0.28 |
| | Major Revision | 1 | 0.54 | 0.92 | 1.00 | 0.54 |
| | Reject | 0 | 0.18 | -0.46 | 0.00 | 0.18 |
| 1.05 | Accept/Minor Revision | 0 | 0.28 | -0.63 | 0.00 | 0.28 |
| | Major Revision | 1 | 0.54 | 0.92 | 1.00 | 0.54 |
| | Reject | 0 | 0.18 | -0.46 | 0.00 | 0.18 |
| 1.08 | Accept/Minor Revision | 0 | 0.29 | -0.63 | 0.00 | 0.29 |
| | Major Revision | 1 | 0.54 | 0.92 | 1.00 | 0.54 |
| | Reject | 0 | 0.17 | -0.46 | 0.00 | 0.17 |
| 1.12 | Accept/Minor Revision | 1 | 0.29 | 1.56 | 1.00 | 0.29 |
| | Major Revision | 0 | 0.54 | -1.08 | 0.00 | 0.54 |
| | Reject | 0 | 0.17 | -0.45 | 0.00 | 0.17 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Publishability Measure | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| 1.13 | Accept/Minor Revision | 0 | 0.29 | -0.64 | 0.00 | 0.29 |
| | Major Revision | 0 | 0.54 | -1.08 | 0.00 | 0.54 |
| | Reject | 1 | 0.17 | 2.22 | 1.00 | 0.17 |
| 1.14 | Accept/Minor Revision | 0 | 0.29 | -0.64 | 0.00 | 0.29 |
| | Major Revision | 1 | 0.54 | 0.92 | 1.00 | 0.54 |
| | Reject | 0 | 0.17 | -0.45 | 0.00 | 0.17 |
| 1.15 | Accept/Minor Revision | 0 | 0.29 | -0.65 | 0.00 | 0.29 |
| | Major Revision | 1 | 0.54 | 0.92 | 1.00 | 0.54 |
| | Reject | 0 | 0.17 | -0.45 | 0.00 | 0.17 |
| 1.16 | Accept/Minor Revision | 0 | 0.30 | -0.65 | 0.00 | 0.30 |
| | Major Revision | 1 | 0.54 | 0.92 | 1.00 | 0.54 |
| | Reject | 0 | 0.17 | -0.45 | 0.00 | 0.17 |
| 1.18 | Accept/Minor Revision | 1 | 0.30 | 1.54 | 1.00 | 0.30 |
| | Major Revision | 0 | 0.54 | -1.08 | 0.00 | 0.54 |
| | Reject | 0 | 0.16 | -0.44 | 0.00 | 0.16 |
| 1.22 | Accept/Minor Revision | 1 | 0.30 | 1.52 | 1.00 | 0.30 |
| | Major Revision | 0 | 0.54 | -1.08 | 0.00 | 0.54 |
| | Reject | 0 | 0.16 | -0.44 | 0.00 | 0.16 |
| 1.25 | Accept/Minor Revision | 1 | 0.31 | 1.51 | 1.00 | 0.31 |
| | Major Revision | 0 | 0.54 | -1.08 | 0.00 | 0.54 |
| | Reject | 0 | 0.16 | -0.43 | 0.00 | 0.16 |
| 1.26 | Accept/Minor Revision | 1 | 0.31 | 1.51 | 1.00 | 0.31 |
| | Major Revision | 0 | 0.54 | -1.08 | 0.00 | 0.54 |
| | Reject | 0 | 0.16 | -0.43 | 0.00 | 0.16 |
| 1.34 | Accept/Minor Revision | 1 | 0.63 | 0.56 | 0.50 | 0.32 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Publishability Measure | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| | Major Revision | 1 | 1.07 | -0.10 | 0.50 | 0.54 |
| | Reject | 0 | 0.30 | -0.59 | 0.00 | 0.15 |
| 1.35 | Accept/Minor Revision | 0 | 1.26 | -1.36 | 0.00 | 0.32 |
| | Major Revision | 4 | 2.14 | 1.86 | 1.00 | 0.54 |
| | Reject | 0 | 0.59 | -0.83 | 0.00 | 0.15 |
| 1.36 | Accept/Minor Revision | 1 | 0.32 | 1.47 | 1.00 | 0.32 |
| | Major Revision | 0 | 0.54 | -1.07 | 0.00 | 0.54 |
| | Reject | 0 | 0.15 | -0.42 | 0.00 | 0.15 |
| 1.39 | Accept/Minor Revision | 0 | 0.32 | -0.69 | 0.00 | 0.32 |
| | Major Revision | 1 | 0.53 | 0.93 | 1.00 | 0.53 |
| | Reject | 0 | 0.14 | -0.41 | 0.00 | 0.14 |
| 1.49 | Accept/Minor Revision | 0 | 0.33 | -0.70 | 0.00 | 0.33 |
| | Major Revision | 1 | 0.53 | 0.94 | 1.00 | 0.53 |
| | Reject | 0 | 0.14 | -0.40 | 0.00 | 0.14 |
| 1.51 | Accept/Minor Revision | 1 | 0.33 | 1.41 | 1.00 | 0.33 |
| | Major Revision | 0 | 0.53 | -1.06 | 0.00 | 0.53 |
| | Reject | 0 | 0.13 | -0.39 | 0.00 | 0.13 |
| 1.52 | Accept/Minor Revision | 2 | 2.35 | -0.28 | 0.29 | 0.34 |
| | Major Revision | 3 | 3.72 | -0.54 | 0.43 | 0.53 |
| | Reject | 2 | 0.94 | 1.18 | 0.29 | 0.13 |
| 1.62 | Accept/Minor Revision | 1 | 0.35 | 1.37 | 1.00 | 0.35 |
| | Major Revision | 0 | 0.53 | -1.06 | 0.00 | 0.53 |
| | Reject | 0 | 0.13 | -0.38 | 0.00 | 0.13 |
| 1.63 | Accept/Minor Revision | 1 | 0.35 | 1.37 | 1.00 | 0.35 |
| | Major Revision | 0 | 0.53 | -1.06 | 0.00 | 0.53 |
| | Reject | 0 | 0.13 | -0.38 | 0.00 | 0.13 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Publishability Measure | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| 1.64 | Accept/Minor Revision | 0 | 0.35 | -0.73 | 0.00 | 0.35 |
| | Major Revision | 1 | 0.53 | 0.95 | 1.00 | 0.53 |
| | Reject | 0 | 0.12 | -0.38 | 0.00 | 0.12 |
| 1.65 | Accept/Minor Revision | 1 | 1.05 | -0.06 | 0.33 | 0.35 |
| | Major Revision | 1 | 1.58 | -0.67 | 0.33 | 0.53 |
| | Reject | 1 | 0.37 | 1.10 | 0.33 | 0.12 |
| 1.76 | Accept/Minor Revision | 0 | 0.36 | -0.75 | 0.00 | 0.36 |
| | Major Revision | 1 | 0.52 | 0.96 | 1.00 | 0.52 |
| | Reject | 0 | 0.12 | -0.36 | 0.00 | 0.12 |
| 1.83 | Accept/Minor Revision | 0 | 0.37 | -0.77 | 0.00 | 0.37 |
| | Major Revision | 1 | 0.52 | 0.96 | 1.00 | 0.52 |
| | Reject | 0 | 0.11 | -0.35 | 0.00 | 0.11 |
| 1.84 | Accept/Minor Revision | 0 | 0.37 | -0.77 | 0.00 | 0.37 |
| | Major Revision | 1 | 0.52 | 0.96 | 1.00 | 0.52 |
| | Reject | 0 | 0.11 | -0.35 | 0.00 | 0.11 |
| 1.88 | Accept/Minor Revision | 1 | 0.75 | 0.36 | 0.50 | 0.38 |
| | Major Revision | 0 | 1.03 | -1.46 | 0.00 | 0.52 |
| | Reject | 1 | 0.21 | 1.80 | 0.50 | 0.11 |
| 1.92 | Accept/Minor Revision | 0 | 0.38 | -0.78 | 0.00 | 0.38 |
| | Major Revision | 1 | 0.51 | 0.97 | 1.00 | 0.51 |
| | Reject | 0 | 0.10 | -0.34 | 0.00 | 0.10 |
| 2.05 | Accept/Minor Revision | 1 | 0.40 | 1.24 | 1.00 | 0.40 |
| | Major Revision | 0 | 0.51 | -1.02 | 0.00 | 0.51 |
| | Reject | 0 | 0.10 | -0.33 | 0.00 | 0.10 |
| 2.22 | Accept/Minor Revision | 0 | 0.42 | -0.84 | 0.00 | 0.42 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Publishability Measure | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| | Major Revision | 1 | 0.50 | 1.00 | 1.00 | 0.50 |
| | Reject | 0 | 0.09 | -0.31 | 0.00 | 0.09 |
| 2.26 | Accept/Minor Revision | 1 | 0.42 | 1.17 | 1.00 | 0.42 |
| | Major Revision | 0 | 0.50 | -0.99 | 0.00 | 0.50 |
| | Reject | 0 | 0.08 | -0.30 | 0.00 | 0.08 |
| 2.27 | Accept/Minor Revision | 1 | 0.42 | 1.17 | 1.00 | 0.42 |
| | Major Revision | 0 | 0.50 | -0.99 | 0.00 | 0.50 |
| | Reject | 0 | 0.08 | -0.30 | 0.00 | 0.08 |
| 2.30 | Accept/Minor Revision | 0 | 0.42 | -0.86 | 0.00 | 0.42 |
| | Major Revision | 0 | 0.49 | -0.99 | 0.00 | 0.49 |
| | Reject | 1 | 0.08 | 3.36 | 1.00 | 0.08 |
| 2.38 | Accept/Minor Revision | 1 | 0.43 | 1.14 | 1.00 | 0.43 |
| | Major Revision | 0 | 0.49 | -0.98 | 0.00 | 0.49 |
| | Reject | 0 | 0.08 | -0.29 | 0.00 | 0.08 |
| 2.43 | Accept/Minor Revision | 1 | 0.44 | 1.13 | 1.00 | 0.44 |
| | Major Revision | 0 | 0.49 | -0.97 | 0.00 | 0.49 |
| | Reject | 0 | 0.07 | -0.28 | 0.00 | 0.07 |
| 2.45 | Accept/Minor Revision | 0 | 0.44 | -0.89 | 0.00 | 0.44 |
| | Major Revision | 1 | 0.48 | 1.03 | 1.00 | 0.48 |
| | Reject | 0 | 0.07 | -0.28 | 0.00 | 0.07 |
| 2.49 | Accept/Minor Revision | 1 | 0.45 | 1.11 | 1.00 | 0.45 |
| | Major Revision | 0 | 0.48 | -0.96 | 0.00 | 0.48 |
| | Reject | 0 | 0.07 | -0.28 | 0.00 | 0.07 |
| 2.51 | Accept/Minor Revision | 1 | 0.90 | 0.14 | 0.50 | 0.45 |
| | Major Revision | 1 | 0.96 | 0.06 | 0.50 | 0.48 |
| | Reject | 0 | 0.14 | -0.39 | 0.00 | 0.07 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Publishability Measure | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| 2.53 | Accept/Minor Revision | 1 | 1.35 | -0.41 | 0.33 | 0.45 |
| | Major Revision | 1 | 1.44 | -0.50 | 0.33 | 0.48 |
| | Reject | 1 | 0.21 | 1.79 | 0.33 | 0.07 |
| 2.55 | Accept/Minor Revision | 0 | 0.45 | -0.91 | 0.00 | 0.45 |
| | Major Revision | 1 | 0.48 | 1.05 | 1.00 | 0.48 |
| | Reject | 0 | 0.07 | -0.27 | 0.00 | 0.07 |
| 2.56 | Accept/Minor Revision | 0 | 0.45 | -0.91 | 0.00 | 0.45 |
| | Major Revision | 1 | 0.48 | 1.05 | 1.00 | 0.48 |
| | Reject | 0 | 0.07 | -0.27 | 0.00 | 0.07 |
| 2.61 | Accept/Minor Revision | 1 | 0.46 | 1.08 | 1.00 | 0.46 |
| | Major Revision | 0 | 0.47 | -0.95 | 0.00 | 0.47 |
| | Reject | 0 | 0.07 | -0.27 | 0.00 | 0.07 |
| 2.66 | Accept/Minor Revision | 0 | 0.47 | -0.93 | 0.00 | 0.47 |
| | Major Revision | 1 | 0.47 | 1.06 | 1.00 | 0.47 |
| | Reject | 0 | 0.06 | -0.26 | 0.00 | 0.06 |
| 2.69 | Accept/Minor Revision | 0 | 0.47 | -0.94 | 0.00 | 0.47 |
| | Major Revision | 0 | 0.47 | -0.94 | 0.00 | 0.47 |
| | Reject | 1 | 0.06 | 3.88 | 1.00 | 0.06 |
| 2.70 | Accept/Minor Revision | 1 | 0.94 | 0.08 | 0.50 | 0.47 |
| | Major Revision | 1 | 0.93 | 0.09 | 0.50 | 0.47 |
| | Reject | 0 | 0.12 | -0.36 | 0.00 | 0.06 |
| 2.78 | Accept/Minor Revision | 0 | 0.48 | -0.96 | 0.00 | 0.48 |
| | Major Revision | 1 | 0.46 | 1.08 | 1.00 | 0.46 |
| | Reject | 0 | 0.06 | -0.25 | 0.00 | 0.06 |
| 2.79 | Accept/Minor Revision | 1 | 0.48 | 1.04 | 1.00 | 0.48 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Publishability Measure | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| | Major Revision | 0 | 0.46 | -0.92 | 0.00 | 0.46 |
| | Reject | 0 | 0.06 | -0.25 | 0.00 | 0.06 |
| 2.99 | Accept/Minor Revision | 1 | 0.50 | 0.99 | 1.00 | 0.50 |
| | Major Revision | 0 | 0.45 | -0.90 | 0.00 | 0.45 |
| | Reject | 0 | 0.05 | -0.23 | 0.00 | 0.05 |
| 3.07 | Accept/Minor Revision | 0 | 0.51 | -1.03 | 0.00 | 0.51 |
| | Major Revision | 1 | 0.44 | 1.13 | 1.00 | 0.44 |
| | Reject | 0 | 0.05 | -0.22 | 0.00 | 0.05 |
| 3.16 | Accept/Minor Revision | 0 | 0.52 | -1.05 | 0.00 | 0.52 |
| | Major Revision | 1 | 0.43 | 1.15 | 1.00 | 0.43 |
| | Reject | 0 | 0.04 | -0.22 | 0.00 | 0.04 |
| 3.17 | Accept/Minor Revision | 1 | 0.52 | 0.95 | 1.00 | 0.52 |
| | Major Revision | 0 | 0.43 | -0.87 | 0.00 | 0.43 |
| | Reject | 0 | 0.04 | -0.22 | 0.00 | 0.04 |
| 3.25 | Accept/Minor Revision | 0 | 0.53 | -1.07 | 0.00 | 0.53 |
| | Major Revision | 1 | 0.43 | 1.16 | 1.00 | 0.43 |
| | Reject | 0 | 0.04 | -0.21 | 0.00 | 0.04 |
| 3.32 | Accept/Minor Revision | 1 | 0.54 | 0.92 | 1.00 | 0.54 |
| | Major Revision | 0 | 0.42 | -0.85 | 0.00 | 0.42 |
| | Reject | 0 | 0.04 | -0.20 | 0.00 | 0.04 |
| 3.39 | Accept/Minor Revision | 0 | 0.55 | -1.10 | 0.00 | 0.55 |
| | Major Revision | 0 | 0.41 | -0.84 | 0.00 | 0.41 |
| | Reject | 1 | 0.04 | 5.05 | 1.00 | 0.04 |
| 3.46 | Accept/Minor Revision | 1 | 0.56 | 0.89 | 1.00 | 0.56 |
| | Major Revision | 0 | 0.41 | -0.83 | 0.00 | 0.41 |
| | Reject | 0 | 0.04 | -0.19 | 0.00 | 0.04 |

| Observed and Predicted Frequencies | | | | | | |
|---|---|---|---|---|---|---|
| | | Frequency | | | Percentage | |
| Publishability Measure | | Observed | Predicted | Pearson Residual | Observed | Predicted |
| 3.55 | Accept/Minor Revision | 1 | 0.57 | 0.88 | 1.00 | 0.57 |
| | Major Revision | 0 | 0.40 | -0.82 | 0.00 | 0.40 |
| | Reject | 0 | 0.03 | -0.19 | 0.00 | 0.03 |
| 3.65 | Accept/Minor Revision | 1 | 0.58 | 0.86 | 1.00 | 0.58 |
| | Major Revision | 0 | 0.39 | -0.80 | 0.00 | 0.39 |
| | Reject | 0 | 0.03 | -0.18 | 0.00 | 0.03 |
| 4.53 | Accept/Minor Revision | 1 | 0.67 | 0.71 | 1.00 | 0.67 |
| | Major Revision | 0 | 0.32 | -0.68 | 0.00 | 0.32 |
| | Reject | 0 | 0.02 | -0.13 | 0.00 | 0.02 |
| 4.74 | Accept/Minor Revision | 0 | 0.68 | -1.47 | 0.00 | 0.68 |
| | Major Revision | 1 | 0.30 | 1.52 | 1.00 | 0.30 |
| | Reject | 0 | 0.01 | -0.12 | 0.00 | 0.01 |
| 8.77 | Accept/Minor Revision | 0 | 0.92 | -3.34 | 0.00 | 0.92 |
| | Major Revision | 1 | 0.08 | 3.35 | 1.00 | 0.08 |
| | Reject | 0 | 0.00 | -0.02 | 0.00 | 0.00 |